Metagenomes and Metatranscriptomes of Cellulolytic Communities in Anoxic Environments

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Anshul Gupta

April 2017

Acknowledgements

I am extremely grateful to my supervisors Prof Alan McCarthy and Dr Heather Allison for their fantastic supervision as well as for advising and motivating me at the right times. Their ability to suggest solutions for seemingly impossible problems will never cease to amaze me. To Alan in particular, thank you for being patient with me during the write-up. It has been a genuine pleasure to work with you, something that was never in doubt even before I started the PhD. I have thoroughly enjoyed the sampling trips and conferences, as well as the many informal chats about anything and everything, not least football.

I would like to thank Dr Mal Horsburgh and Dr Meriel Jones for always being there for a friendly chat and to provide advice or support. I am also thankful to NERC for providing me with an opportunity to undertake my PhD studentship. I will always remember Paul Loughnane not only for being the best lab technician, but also for being the person who taught me the fine arts of molecular biology (in between chats about how my Friday night at Krazyhouse was) all the way back in the summer before my honours year during my undergraduate degree!

I have had the fortune of working alongside some brilliant postdocs: Dr Dave Rooks, Dr Marta Veses Garcia, Dr Bruno Manso and Dr Disa Hammarlof to name some. Working in lab H was a pleasure as they not only taught me research skills, but also that Pub Friday will always be there regardless of what happens in lab during the week. Thank you to Dr James Houghton who taught me some tricks of bioinformatics when I was struggling with it. Special thanks must also go to Dr James McDonald for sharing some of his considerable knowledge on cellulose degradation and for being a friendly face at conferences, as well as to our friends and colleagues at Bangor University and University of Huddersfield (even though I am still a little unhappy with them for introducing me to the world of pain that is anaerobic hyperalkaline sediments!).

I have been fortunate enough to find the company of like-minded people in lab H and the institute, and have managed to forge some friendships that will hopefully last forever. Seanyg123, Reiss, Stewie, Brae-know, Babs, Hilly, Dani, Amy and Weeg: thank you for always being willing to listen to my numerous rants when the overlords of science were not appeased, and for joining me for a drink or ten at my second home, the AJ (first home being the lab). Not to mention the friends I've lived with and played football with. My time in Liverpool has been particularly enjoyable, and it wouldn't have been the same without each and every one of you. It is also impossible not to mention the loons back home who ensured that I remained (relatively) sane when times were testing, especially over the writing up period.

Last but not the least, huge thanks to my parents and my brother. I will not be where I am without their constant unconditional support, both financial and emotional. Thank you for also being extremely patient with me, particularly over the final months, as I sulked away in a corner trying to finish the thesis.

Abstract

Anaerobic microbial communities that degrade cellulose are structurally and metabolically complex, and studies addressing the abundance and relative activity of their constituent microbial groups have mainly focused on the herbivore gut. As a consequence, relatively unexplored anoxic environments such as landfill sites can be viewed as a potential reservoir of novel cellulolytic agents. High-throughput sequencing of metagenomes and metatranscriptomes enables the analysis of these diverse microbial communities rich in uncultivable species, whilst overcoming the bias associated with the use of amplicon clone libraries. This study surveyed cellulolytic microbes present both in landfill leachate as well as those in association with cotton cellulose 'baits' maintained in leachate microcosms. PCR detection of anaerobic chytrid fungi, some of the most potent cellulose degraders in the microbial world, from certain leachate and cotton samples was followed by an unsuccessful attempt at culturing these microbes from landfill leachate. Identity of microbes responsible for amplification of 18S rRNA genes from leachate was queried by sequencing of a subset of a clone library, where ciliated protozoa of Class Armophorea accounted for 78% of all sequences. QIIME analysis of previously 454 pyrosequenced 18S rRNA gene amplicons from leachate and cotton incubated in leachate however, pointed towards dominance of datasets by anaerobic fungal sequences (>99% and 94% of all sequences, respectively).

A metagenome, a metatranscriptome and a MessageAmp II aRNA amplification kit (Invitrogen) amplified metatranscriptome derived from cotton incubated in landfill leachate were investigated following Illumina MiSeq sequencing. Taxonomic profiling using MG-RAST indicated that the datasets were dominated by Bacteria, whereas the percentage of reads assigned to Archaea increased from 6.6% to 14.2%, and those assigned to Eukaryota increased from 1% to 7.6% in the metatranscriptome compared to the metagenome. Proteobacteria, Firmicutes and Euryarchaeota were the predominant phyla in the landfill cellulose metatranscriptome, where unclassified sequences comprised 11% of the total sequences compared to only 1.2% in the metagenome, implying that these microbes are more active in the environment under survey than DNA analysis alone would suggest. No outstanding differences were evident in the taxonomic and functional comparison of the amplified and the non-amplified metatranscriptomes. HMMER alignment of 2.34×10^6 million ORFs predicted in the metatranscriptome revealed that 1,724 matched glycosyl hydrolases involved in polysaccharide degradation, with 114 sequences returning no matches in the NCBI nr database and hence represent potentially novel GH enzymes.

A fosmid library was generated from parallel high molecular weight DNA samples and preliminary expression screening was performed for mining of cellulase-encoding genes. Congo red screening of a second metagenomic fosmid library, generated from cotton incubated in lake sediment, yielded two distinct clones expressing endoglucanases. Sequencing and assembly of one of the clones led to the identification of two genes encoding a GH10 and a GH5 domain.

A method was developed for extraction of high quality total RNA, whilst maintaining sufficient yield, from mixed microbial communities found in anoxic hyperalkaline environments to allow for metatranscriptome-mediated discovery of glycosyl hydrolases. Apart from improving our understanding of cellulose degradation in anoxic environments, novel cellulases have commercial significance in biomass processing including production of second-generation biofuels.

Table of Contents

Abbreviations

BD	Bromborough Dock landfill site
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
BM	Bidston Moss landfill site
bp	Base pair
CAZy	Carbohydrate active enzymes
cDNA	Complementary DNA
CMC	Carboxy methyl cellulose
CTAB	Cetyltrimethylammonium bromide
DEPC	Diethylpyrocarbonate
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
dscDNA	Double-stranded complementary DNA
EDTA	Ethylenediaminetetraacetic
GH	Glycoside hydrolase
HMM	Hidden Markov Model
HMW	High molecular weight
LB	Luria Bertani medium
LSU	Large Subunit
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Information
NR	Non-redundant (i.e. Non-redundant database)
ORF	Open Reading Frame
OTU	Operational Taxanomic Unit
PBS	Phosphate-buffered saline
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
PFGE	Pulse-Field Gel Electrophoresis
PolyA	Poly Adenine
qPCR	Quantitative PCR

RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
sddH_20	Sterile double distilled H_20
SDS	Sodium Dodecyl Sulphate
ssDNA	Single stranded DNA
SSU	Small Subunit
SET	Sucrose-EDTA-Tris
TAE	Tris-acetate-EDTA
TBE	Tris-borate-EDTA
TE	Tris-EDTA
T_m	Melting Temperature
Tris	Tris(hydroxymethyl)aminomethane
v/v	Volume/Volume
w/v	Weight/Volume

IUPAC Degenerate Base Symbols

A	**A**denosine
C	**C**ytidine
G	**G**uanosine
T	**T**hymidine
U	**U**racil
W	**W**eak (A or T)
S	**S**trong (G or C)
M	A**m**ino (A or C)
K	**K**eto (G or T)
R	Pu**r**ine (A or G)
Y	P**y**rimidine (C or T)
B	Not A
D	Not C
H	Not G
V	Not T
N	Any base (or unspecified base)

Chapter 1

Introduction

Microorganisms play an integral role in governing the transformation of energy and mediating the various biogeochemical processes on Earth, by way of forming complex communities consisting of numerous distinct microbes with fundamentally diverse metabolic capabilities. Cellulose is the most abundant organic polymer on Earth, and microbes that degrade it display phylogenetic divergence as well as widespread global distribution. There is therefore concerted interest in not only understanding the immense role of cellulolytic microorganisms in nutrient cycling, but also exploiting their untapped genetic potential for various industrial applications. Advances in high throughput sequencing coupled with an enhanced repertoire of bioinformatic tools allow for pertinent biological conclusions to be drawn more accurately from molecular microbial ecology studies. Consequently, undertaking a comprehensive 'omics'-mediated survey of highly heterogeneous microbial communities in unexplored anoxic environments such as landfill sites that are potentially rich in novel cellulolytic agents is essential.

1.1 Molecular microbial ecology

Molecular microbial ecology refers to the study of the structure and function of microbial communities, as well as the complex interactions between both the microorganisms that comprise them and their environment. It has been estimated that up to 99% of microbes in a given ecosystem cannot be cultured using standard techniques (Amann *et al*., 1995; Pace, 1997; Torsvik & Ovreas, 2002; Curtis & Sloan, 2004; Mendes *et al.*, 2013; Hill *et al*., 2017), the main reason being our inability to precisely duplicate the growth conditions they have become adapted to within their interdependent, heterogeneous communities. It is also widely

acknowledged that culture-dependent methods bias our understanding of microbial diversity and hinder true appreciation of the composition of these microbial communities (Woese, 1996). Investigations into the activity and ecological role of a microbe in its natural environment are best performed within the context of said environment, and should take precedence over culture-based procedures. The influential work of Carl Woese and Norman Pace, centred on using molecular phylogeny to determine the microbial diversity in a given sample, coupled with progress in molecular biological techniques led to the use of polymerase chain reaction (PCR) for amplification and subsequent comparison of the DNA sequence of the small subunit rRNA (SSU rRNA) to characterise microbial populations (Woese, 1987; Pace *et al*., 1986; Olsen *et al*., 1986).

1.1.1 Culture-independent analysis of microbial communities: techniques and approaches

Modern molecular techniques were developed in order to overcome the limitations imposed by pre-existing culture-dependent laboratory procedures. However, no technique is perfect and acknowledging their drawbacks is essential for experimental design and for drawing meaningful conclusions from experiments conducted. The relative merits and demerits of some of these culture-independent techniques are discussed here.

1.1.1.1 PCR

PCR is the most commonly used technique for molecular analysis of microbial communities in their natural environment, and it relies on the extraction of nucleic acids as the primary step. However, inherent differences in the amplification efficiency of templates gives rise to PCR amplification bias (Suzuki & Giovannoni, 1996; Polz & Cavanaugh, 1998). This phenomenon can primarily be attributed to differences in primer-template binding energy (Ishii & Fukui, 2001) and differences in DNA polymerase binding during initiation of polymerisation

(Aird *et al*., 2011; Dabney & Meyer, 2012), and can skew estimates of microbial diversity in multi-template environmental samples.

1.1.1.2 Cloning

PCR-generated amplicons of the 16S and 18S rRNA genes from Archaea, Bacteria and Eukarya have been widely used to generate clone libraries for the determination of the phylogenetic composition of communities. Cloning has been used in conjunction with Sanger sequencing to survey the plant biomass degrading capability of termite hindgut (Warnecke *et al*., 2007), while several examples can be found illustrating the role of cloning in phylogenetic analysis of environmental samples (Dawson & Pace, 2002; de Menezes *et al*., 2008). However, focus is now switching towards high throughput sequencing of amplicons due to the inherent biases and inaccuracies introduced by PCR amplification (Hong *et al*., 2009).

Expression library screening of cloned environmental DNA has been used extensively for novel enzyme discovery, and high throughput studies incorporating metagenomic DNA from whole microbial communities are now common (Rooks *et al*., 2012). The definite advantage associated with shotgun cloning is that it produces actual analysable inserts rather than just output data in the form of sequence reads.

1.1.1.3 Molecular fingerprinting techniques

DNA fingerprinting methods such as Denaturing Gradient Gel Electrophoresis (DGGE), Temperature Gradient Gel Electrophoresis (TGGE) and Restriction Fragment Length Polymorphism (RFLP) have been useful in observing microbial community response to changing environmental conditions (Deng *et al*., 2008), as community DNA extractions can be examined for shifts in structure and composition. PCR amplification of the 16S rRNA gene or cDNA generated from the reverse transcribed 16S rRNA from the microbial community is

followed by electrophoresis of the amplicons on a gel where a gradient is created using either temperature (TGGE) or a denaturing agent, usually acrylamide (DGGE). The resulting pattern of bands is obtained due to partial dissociation of the amplicons, and has been used in environmental microbiology (e.g. Gray *et al*., 2003; Nicol *et al*., 2007; Shin *et al*., 2008).

Terminal-RFLP (T-RFLP) generates specific banding patterns that can be visualised through the detection of fluorescent dye, after fluorescently labelled primers are used to generate PCR amplicons that are digested with a restriction enzyme. However, molecular fingerprinting methods are practically redundant as banding patterns on a gel are only partially informative, and as such, amplicon sequencing has emerged as a superior alternative with the advent of high throughput sequencing. Despite that, T-RFLP is still used in environmental microbiology studies where multiple comparisons of a large number of samples is necessary (e.g. Reider & Frey, 2013; Zumsteg *et al*., 2012; Ying *et al*., 2013).

1.1.1.4 Quantitative PCR

Quantitative PCR (qPCR) has been a mainstay of studies aiming to provide insights into microbial communities, as it is a powerful technique for the estimation of the copy number of particular genes. qPCR's major advantage over conventional PCR is that it details the relative abundance of target taxa rather than simply report on their presence. For example, McDonald *et al*. (2008; 2009) used qPCR and molecular cloning to investigate the abundance of the cellulolytic bacterial genus *Fibrobacter* in landfill leachate and freshwater lake sediment, and demonstrated the presence of novel lineages in those anoxic environments. The procedure in itself is, however, technically challenging and quality data must be generated through rigorous experimental design, with meticulously controlled and replicated experiments with robust internal standards and target gene validation a necessity (Marx, 2013). Data interpretation has also been reported to be a potential issue with qPCR, as poor quality RNA

in particular has been reported to cause high levels of variability whilst determining absolute numbers of transcripts between technical replicates (Smith *et al*., 2006). It is therefore, recommended that qPCR data should be expressed as relative abundance of transcripts or gene copy number as that will aid comparisons between assays.

1.1.1.5 Molecular probing

Estimating the abundance of specific phylogenetic groups through the design of taxon-specific oligonucleotide sequences can be achieved by utilising information gathered from clone library content, qPCR or high throughput methods. The major drawback of this technique is that prior knowledge of community composition is a prerequisite (Frickmann *et al*., 2017). Fluorescent *in situ* Hybridisation (FISH), however, can be extremely useful as a probing technique as it facilitates visualising the actual location of a particular group of microbes within a target environment, a feat not attainable using amplification, cloning or sequencing (Frickmann *et al*., 2017).

1.1.2 High throughput culture-independent techniques

Metagenomics and metatranscriptomics are the two most commonly used high throughput 'omics' techniques in environmental microbiology studies. Metagenomics refers to the study of the collective genomic material present within all the microbes in an environment, while metatranscriptomics refers to the study of the collective gene expression profile within a group of interacting microbes in an environment. The first step in all metagenomic studies is the extraction of DNA from all the representatives of a microbial community, whereas metatranscriptomics relies on the extraction of total RNA from a similar sample. High throughput sequencing of DNA, SSU rRNA and mRNA enables the analyses of these diverse microbial communities rich in uncultivable species, whilst overcoming the bias associated with

the use of amplicon clone libraries to determine diversity and relative abundances (Shi *et al*., 2011), as well as enabling via metatranscriptomics an appreciation of *in situ* microbial gene expression and activity.

Recent technological advancements have resulted in multiple sequencing platforms that offer a highly appealing alternative to Sanger sequencing (Sanger *et al*., 1977). The most crucial difference between next generation sequencing and traditional Sanger sequencing is the throughput, as we can now generate 10^4-10^7 times more sequences per sequencing run (Mardis, 2017). 454 pyrosequencing (Roche) was the preferred platform until recently (Frias-Lopez *et al*., 2008; Luo *et al*., 2012) as it generated sequence reads of up to 1000 bp, which aided assembly and taxonomic annotation. However, Illumina sequencing is now the preferred platform for application in metagenomics due to its much lower error rate associated with A-T rich homopolymer sequences, along with the cost effectiveness brought about by improvements in read length and higher throughput than pyrosequencing (Luo *et al*., 2012). A number of sequencing platforms are currently available, offering varying levels of sequence coverage and depth, and a summary is presented in Table 1.1.

Table 1.1. Summary of currently available and commonly used next generation sequencing technology platforms (adapted from Goodwin *et al*., 2016). This list is not exhaustive.

Platform	**No. of reads (x10^6)**	**Read length (bp)**
ABI SOLiD 5500 xl	~ 1,400*	75
454 GS Junior+	~ 0.1	700
454 GS FLX Titanium XL+	~ 1	700
Illumina HiSeq2500 v2 Rapid run	600	250+250^
Illumina HiSeq2500 v4	4,000	125+125^
Illumina HiSeq X	3,000	150+150^
Illumina MiniSeq High output	~ 50	150+150^
Illumina MiSeq v3	~ 50	300+300^
Illumina NextSeq 500/550 High output	800	150+150^
Ion PGM 318	4-5.5*	400
Ion Proton	60-80*	200
Ion S5 540	60-80*	200
Pacific Biosciences RS II	~ 0.05*	~ 20 kb
Pacific Biosciences Sequel	~ 0.35*	8-12 kb
Oxford Nanopore MK1 MinION	0.1§	Up to 200 kb§

* manufacturer's data
^ paired-end sequencing
§ Ip *et al.* (2015)

1.1.2.1 Metagenomics

Most metagenomic studies involving high throughput sequencing adopt either an amplicon based or a shotgun approach. Amplification of the 16S rRNA gene (Archaea and Bacteria), 18S rRNA gene (Eukaryota) or the Internal Transcribed Spacer (ITS) region followed by subsequent sequencing provides data facilitating comprehensive taxonomic analysis of microbial species. Alternatively, a shotgun library of sequences can be produced by virtue of direct community DNA extraction from an environmental sample. Such an approach yields data rich in phylogenetic information and offers an insight into the metabolic potential of the population, whilst also aiding direct gene mining for biotechnological applications (Scholz *et al*., 2012). The community DNA can also be cloned into a heterologous host (usually *E. coli*) to be expressed as proteins to facilitate gene mining from previously unexplored environments, albeit with the limitation that host specific expression is required.

Functional metagenomics aims not only to carry out an in-depth survey of the microbial diversity in an environmental sample, but also to assign a role to the catalogue of genes produced from members of the microbial community (De Filippo *et al*., 2012). Discovery of novel genes and enzymes from environmental metagenomes is greatly facilitated by functional metagenomics approaches, and several studies highlight the efficacy of such methods for discovery of enzymes for plant biomass degradation. While Brulc *et al*. (2009) used 454 pyrosequencing to determine the glycoside hydrolase profile of microbial communities responsible for fibre degradation in the rumen, Hess *et al*. (2011) used paired-end Illumina sequencing to generate ~1.5 billion reads from microbial communities degrading switchgrass in the rumen. Open Reading Frame (ORF) prediction followed by subsequent PCR amplification and molecular cloning enabled the overexpression of enzymes of interest, whilst the supreme depth of

sequencing also allowed for assembly (60%-93%) of draft genomes of uncultivated microbes.

1.1.2.2 Functional metagenomics using fosmids

Metagenomic approaches relying on community DNA sequencing have a vital role to play in microbial ecology, particularly given the continuing technological advances in high throughput sequencing and data analyses. Owing to the fact that a large proportion of newly discovered environmental sequences show no significant similarity to previously annotated sequences, assigning biochemical functions to them however remains challenging. The preparation of metagenomic clone libraries allows for functional screening of environmental DNA to provide evidence for function of previously unknown genes, with the added benefit of requiring no previous knowledge of those sequences whilst also bypassing the bias associated with amplification techniques (Rooks *et al*., 2012).

While plasmids have been used extensively for the cloning and subsequent overexpression of genes of interest, cloning into cosmids, fosmids and Bacterial Artificial Chromosomes (BACs) have also been utilised for expression screening of metagenomic DNA for the mining of novel genes and their proteins. The preparation of high molecular weight DNA (30-50 kb) clone libraries using fosmids as vectors can prove to be an effective method for gene mining in metagenomic studies, as this allows for whole genes and gene clusters to be cloned randomly from complex environmental microbial communities, facilitating screening of the insert DNA for novel enzymes of industrial and biotechnological significance (Singh & Macdonald, 2010; Ekkers *et al*., 2012). Furthermore, sequencing of such complex gene clusters can disclose the role of accessory or regulatory components related to the function of the enzymes identified

during screening, and as such, fosmids could also prove to be useful tools for investigating the genomic context of related genes during metagenomic studies (Palackal *et al*., 2007). Screening of metagenomic libraries is usually performed using either a functional approach or a sequence-based approach (Schloss & Handelsman, 2003; Kakirde *et al*., 2010), where the focus is either on the detection of an expressed enzyme or on the detection of a target gene sequence by PCR or molecular hybridisation, respectively.

Metagenomic libraries have been constructed from a broad range of environments in order to survey the genetic potential of the microbial communities present, as well as for gene mining, leading to the discovery of cellulases, xylanases, amylases, chitinases, lipases, proteases and antibiotics (Ekkers *et al*., 2012). Some of these environments include soil and sediment (Brennerova *et al*., 2009; Parsley *et al*., 2010; Berini *et al*., 2017), freshwater and marine environments (Wexler *et al*., 2005; Martin-Cuadrado *et al*., 2007), guts of animals (Li *et al*., 2008; Wang *et al*., 2011; Song *et al*., 2017), Arctic and glacial ice (Jeon *et al*., 2009; Simon *et al*., 2009) as well as hyperthermal pond (Rhee *et al*., 2005).

1.1.2.3 Metatranscriptomics

High throughput sequencing of RNA libraries is usually referred to as RNA-seq, which can be misleading as current sequencing platforms lack the capability to directly sequence RNA and hence reverse transcription into a constitutive cDNA library is necessary as the preliminary step. rRNA depletion is necessary in studies where the enzyme-coding expression profile of a microbial community is under investigation as mRNA makes up only 1-10% of the total RNA in a given population, and sequencing of total RNA would imply that only a small proportion of the sequencing throughput

would be used efficiently (He *et al*., 2010). Commonly used rRNA removal methods include subtractive hybridisation, where pre-designed probes hybridise to known microbial SSU and LSU rRNA to facilitate physical removal, and treatment with 5'-monophosphate-dependent exonuclease enzyme (Mader *et al*., 2011).

Failing to remove rRNA sequences can lead to misannotation of putative mRNA reads when analysis is performed for matching protein homologs (Tripp *et al*., 2011). However, studies investigating both gene expression profile and taxonomic make-up of microbial communities have performed high throughput sequencing of total RNA, followed by separate analysis of mRNA and rRNA sequences (Urich *et al*., 2008; Radax *et al*., 2012). A functional metatranscriptomic methodology has previously been used for identification of genes encoding cellulolytic enzymes. Takasaki *et al*. (2013) sequenced mRNA from a soil microbial community that was enriched using cellulose, followed by querying putative ORFs against NCBI nr database to identify glycoside hydrolases involved in cellulose hydrolysis.

While protein synthesis can be regulated by post-transcriptional and post-translational gene expression, microbes commonly adapt to changing environmental conditions by regulating gene expression at the transcriptional level. This makes the study of immediate regulatory responses to altering conditions through metatranscriptomics an attractive alternative to the technologically challenging science of metaproteomics (Moran, 2009). Metagenomics and metatranscriptomics allow us to compare and contrast the genetic potential with the *in situ* functional activity in a complex community of microbes, and have the potential to be integral tools in molecular microbial ecology studies aimed at interpreting biology at the aggregate level. As such, rapid technological advances in high throughput sequencing make

transcriptomics an exciting alternative to traditional gene expression profiling of microbes using microarray analysis.

The DeLong research group has published papers detailing metatranscriptomic approaches used for community structure and gene expression analyses of marine microbial communities (Frias-Lopez *et al.*, 2008; Shi *et al*., 2009; Stewart *et al*., 2010; Shi *et al*., 2011). This is significant as the phenomenon of a 'rare biosphere' was observed i.e. most natural environments, including wastewater treatment and landfill sites, are dominated by a handful of species while a large amount of microbial diversity exists at a low abundance (first reported by Sogin *et al*., 2006). Furthermore, a large proportion of the sequences in metatranscriptomes derived from environmental microbial communities either constitute a hypothetical protein of unknown function, or tend to be missing from the NCBI nr database (Frias-Lopez *et al.*, 2008), a problem greatly enhanced in metaviromes, where the vast majority of sequences have no match.

Metatranscriptomic overview of microbial communities exists from a number of environments yet very few studies have compared them to metagenomes, contemporaneously. To date, no studies have focussed on the taxonomic and functional profiling of the cellulose degrading microbial community in landfill leachate. So, the application of metagenomics and metatranscriptomics to the discovery of novel cellulases in the anaerobic and highly active microbial community of landfill sites was a driver of the research presented in this thesis.

1.2 Cellulose degradation: microbiology and industrial significance

The process of photosynthesis is central to life on Earth, and gives rise to a substantial amount of plant biomass. Cellulose is the primary structural component of this biomass (Lynd *et al*., 2002), whilst also being present within certain bacteria, fungi,

algae and animals (O'Sullivan, 1997). Therefore, the global carbon cycle relies fundamentally on the degradative action of microbes that mineralise carbon, and as such are responsible in part for the largest material flow in the biosphere (Lynd *et al*., 2002).

1.2.1 Cellulose

Cellulose is the most abundant naturally occurring carbon polymer on earth. It is a straight-chained, unbranched homopolymer consisting of glucose molecules joined together by β-1,4 glycosidic bonds and is the main structural component of higher plant cell walls (Fig. 1.1), representing approximately 35-50% of plant dry weight. The monomeric unit of cellulose is called cellobiose, and is a disaccharide consisting of two glucose units that have been rotated by 180º relative to each other (Beguin & Aubert, 1994). Due to the variation in the degree of polymerisation between primary and secondary plant cell wall, the length of cellulose chains can vary between 500 and 14,000 glucose moieties, with primary cell wall usually consisting of much fewer residues than secondary cell wall (Ljungdahl & Eriksson, 1985).

Extensive hydrogen bonding and van der Waal's forces between individual molecules leads to the arrangement of cellulose chains into insoluble microfibrils, which in turn form cellulose fibres. Parallel orientation of cellulose chains allows for the formation of highly ordered crystalline regions interspersed with more disordered amorphous domains. Such crystalline structure makes cellulosic biomass highly resistant to degradation, whilst also enabling plants to resist mechanical stress and turgor pressure. Depending upon its origin, the degree of crystallinity in native cellulose can vary between 60 and 99%, with the secondary walls of cotton seed hairs being nearly 100% crystalline cellulose (Marchessault & Sundararajan, 1983).

However, such pure forms of cellulose are rarely found in nature, as it is almost always associated with other plant substances, including hemicellulose, lignin and pectin. Hemicelluloses comprise of a diverse group of linear and branched heteropolymers of D-xylose, L-arabinose, D-mannose, D-glucose, D-galactose and D-glucuronic acid (Leschine, 1995). The most abundant hemicelluloses are glucomannans and xylans, the latter consisting of acetyl, methylglucuronyl and arabinofuranosyl side chains. These side chains can form hemicellulose-lignin cross-linkages through esterification by the aromatic acids associated with lignin, which include *p*-hydroxyphenyls, *p*-guaicyls, *p*-coumaryls and *p*-syringyls (Buranov & Mazza, 2008; Chundawat *et al*., 2011). Lignin is a recalcitrant heterogeneous polymer, which consists of a highly branched structure of variable composition. Covalent bond formation between lignin and hemicellulose in the plant cell wall leads to cellulose microfibrils being embedded within a matrix of polymers. A high degree of lignification leads to the formation of rigid, woody tissue, whilst also acting as a physical barrier to restrict the access of cellulases and hemicellulases to cellulose and hemicellulose, respectively (Mooney *et al*., 1998; Laureano-Perez *et al*., 2005). Consequently, cellulosic biomass is highly resistant to enzymatic degradation.

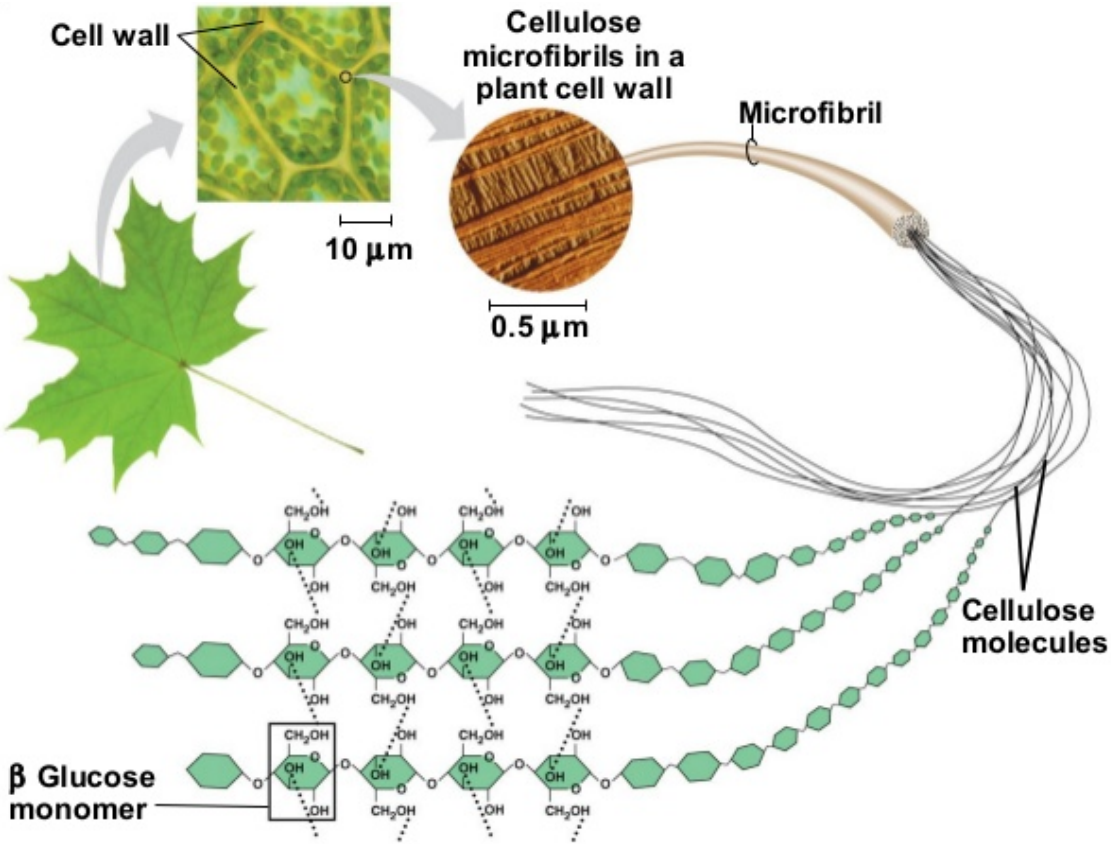


Figure 1.1. Illustration of cellulose structure consisting of glucose molecules joined together by β-1,4 glycosidic bonds, as well as arrangement of cellulose chains to form microfibrils (from Barley & Fitzpatrick, 2011).

1.2.2 Industrial importance of cellulases

Apart from improving our understanding of cellulose degradation in anoxic environments such as landfill sites, discovery and characterisation of novel cellulases has commercial applications, as these enzymes are considered to be the third most significant industrial enzymes on the global market after amylases and proteases (Sajith *et al*., 2016). Examples of uses include in agriculture, waste management and various industries, including for pulping and deinking in paper and pulp industry, for bio-polishing and biostoning in textile industry, for fabric softening in laundry and detergent industry, for extraction and enhancing digestibility in food and feed industry, as well as in brewing and chitosan production (Kuhad *et al*., 2011; Sajith *et al*., 2016). The substantial rise in global energy usage and the escalated depletion of fossil fuel reserves, however, makes the production of renewable energy resources like bioethanol the most important industrial application of cellulases. Issues with climate change further highlight the growing need for clean, carbon-neutral fuels, leading to extensive research in this field (Ragauskas *et al*., 2006; Schubert, 2006). The most common method relies on the generation of high-energy bioethanol using renewable plant material such as corn and sugarcane (Olson *et al*., 2012).

Bioethanol has the potential to be a suitable substitute for fossil fuels as it is compatible with conventional petroleum fuels and its combustion produces relatively lower greenhouse gas emissions (Morrison *et al*., 2009). Since the 1980s, the United States has been using fuel-grade ethanol generated from corn as an E10 blend (a blend of gasoline mixed with 10% bioethanol by volume), and more than 96% of gasoline currently sold in the USA contains ethanol (US Department of Energy, 2015). Following the generation of 4.9 billion gallons in 2006, there has been a steady increase in the annual production of ethanol in the USA, and the annual production capacity was

determined to be over 14.9 billion gallons in 2014 (US Energy Information Administration, 2014). However, production of first generation biofuel has limitations, primarily because only parts of the plant that contain fermentable sugars are utilised in first-generation bioethanol production, raising unanswered questions over the long-term sustainability of such a practice.

Patzek & Pimentel (2005) reported that intensive farming of corn fed into bioethanol plants consumes more energy than is generated from the combustion of the bioethanol produced from it. While the energy balance of ethanol currently produced from corn is now positive, an estimated up to 30-fold improvement in the energy balance is possible by using cellulosic corn stover as biomass for bioethanol generation (United States Department of Agriculture, 2016). The Renewable Fuel Standard (RFS) in the USA limits the amount of ethanol generated from starch-based feedstocks to 15 billion gallons per annum to ensure sufficient feedstock availability to meet demands in food and livestock feed, according to the Energy Independence and Security Act of 2007. It has been estimated that cellulosic ethanol could reduce greenhouse gas emissions by up to 86% compared to gasoline in the USA (US Department of Energy, 2015).

Hence, there is an increased demand for the utilisation of lignocellulosic biomass as raw material for the production of second-generation bioethanol, not least to alleviate the current reliance on food sources for the generation of biofuels. This process involves the pre-treatment and hydrolysis of lignocellulosic biomass into single-chain polysaccharides, followed by the fermentation of these sugars into bioethanol (Rubin, 2008). The use of plants that assimilate CO_2 into a C_4 compound following photosynthesis has been advocated as the potential substrate for bioethanol production due to their high productivity (Sanderson, 2006). Examples include

perennial grasses such as *Miscanthus* and switchgrass, as these have the advantage of high biomass density, rapid growth even in poor land with low nutrient and water content as well as the non-requirement of annual replantation after a harvest (Rubin, 2008). Novozymes opened the world's first commercial-scale facility for producton of cellulosic ethanol in Crescentino, Italy, in 2012. They use their Cellic enzyme for hydrolysis of feedstock (including wheat straw, rice straw and energy crops like giant cane) followed by subsequent fermentation of the sugars generated, and have an annual production potential of 75 million litres (Novozymes, accessible at http://www.novozymes.com/en/news/image/crescentino-grand-opening).

Given the recalcitrant nature of lignocellulosic substrate, harsh physiochemical conditions have been used in the initial hydrolysis of lignocellulose in industry (Wolfenden *et al*., 1999). Common examples include steam explosion using sulphuric acid (Shuai *et al*., 2010; Zhu *et al*., 2010), ammonia fibre explosion (Teymouri *et al*., 2005; Wyman *et al*., 2009) and use of alkaline peroxide (Gould, 1984; Gould, 1985) and liquid hot water (Mosier *et al*., 2005; Wyman *et al*., 2009). This adds considerable expense to the process, particularly due to the need for the construction of specific reactors.

Recent advances in expression efficiency of cellulases alone have rendered enzymatic saccharification economically viable (Himmel & Picataggio, 2008; Gusakov, 2011; Hasunuma *et al*., 2013). Traditionally, biological saccharification has been carried out using microbes such as soft-, white- and brown-rot fungi (Schurz, 1978), with the use of enzymes such as cellulases, xylanases and carbohydrate esterases amongst others now popular due to the low energy requirement during hydrolysis. However, the rate of hydrolysis continues to remain low, necessitating the need for the

discovery of novel enzymes, including cellulases, which demonstrate considerably higher rate of depolymerisation.

The isolation of novel cellulases from anoxic environments is important, as known cellulases from anaerobic microorganisms have demonstrated higher specific activity than those produced by aerobic microorganisms (Bayer *et al*., 2008; Munir & Levin, 2016). The energy constraints imposed by the anoxic environment have led to the evolution of highly efficient cellulosome systems, with that of *Clostridium thermocellum* reported to be 50-fold more active in degrading crystalline cellulose than the archetypal cellulase system of the aerobic fungus *Trichoderma reesei* system (Demain *et al*., 2005). Moreover, the potent cellulases complex of *Neocallimastix frontalis* was reported to hydrolyse four times more cotton fibre than *Trichoderma reesei* C30 over a period of 48 hours (Wood *et al*., 1986). Therefore, anaerobically produced cellulases have significant potential to be utilised in industry.

1.2.3 Mechanisms and enzymes involved in cellulose degradation

Despite the fact that the majority of cellulose degradation occurs under aerobic conditions (~85-90%), the definitive amount of cellulose degraded under anaerobic conditions is immense and environmentally significant (Vogels, 1979; Jenkinson *et al*., 1991). The most significant difference in the process of cellulose decomposition between aerobic and anaerobic microbes is that while individual species might be able to achieve it aerobically, a complex microbial community is usually found associated with anaerobic degradation (Leschine, 1995). The various environmental factors that define an anoxic environment play a crucial role in influencing the microbial community found associated with it, making the degradation of cellulose by numerous metabolically diverse microbes a complex process dictated by the physiological and

biochemical interactions of such a community (Leschine, 1995). Consequently, relatively little is known about the microbiology of cellulose decomposition in anoxic environments such as landfill sites.

1.2.3.1 Aerobic cellulose degradation- the free cellulase system

The two major mechanisms employed by cellulolytic microbes are the free cellulase mechanism (Wilson, 2008) and the cell surface-bound cellulases mechanism (Bayer *et al*., 2004; Ransom-Jones *et al*., 2012) (Fig. 1.2 a,b). The cell-free enzyme mechanism is typical of aerobic bacteria and fungi, and is based on the cellulase system of the aerobic fungus *Trichoderma reesei*. Extracellular cellulases released by these microbes carry their own binding molecules (Himmel & Picataggio, 2008). These enzymes act synergistically to facilitate cellulose degradation, particularly in the case of aerobic filamentous fungi and actinomycetes, as these microbes are able to gain access to confined areas by penetrating the cellulose fibres (Lynd *et al*., 2005).

1.2.3.2 Anaerobic cellulose degradation: the cellulosome system

Anaerobic bacteria and chytrid fungi use cell-bound complexes called cellulosomes (Fig. 1.3) to achieve cellulose hydrolysis (Fontes & Gilbert, 2010; Artzi *et al*., 2017). Cellulosomes are large (MDa) multi-protein complexes consisting of discrete enzymes and structural proteins (Lynd *et al*., 2005; Bayer *et al*., 2004; Smith & Bayer, 2013). Structural components of cellulosomes include scaffoldin proteins, which are found attached to the bacterial cell wall. These scaffoldins in turn carry cohesin proteins that attach to the dockerin domains of the various cellulases and hemicellulases produced by the microbes. Scaffoldins also carry a carbohydrate binding

module (CBM) that attaches to the cellulosic substrate, and this facilitates the synergistic interactions between various classes of enzymes that are necessary to cause cellulose depolymerisation (Fontes & Gilbert, 2010; Gilbert *et al*., 2013; Smith & Bayer, 2013; Artzi *et al*., 2017).

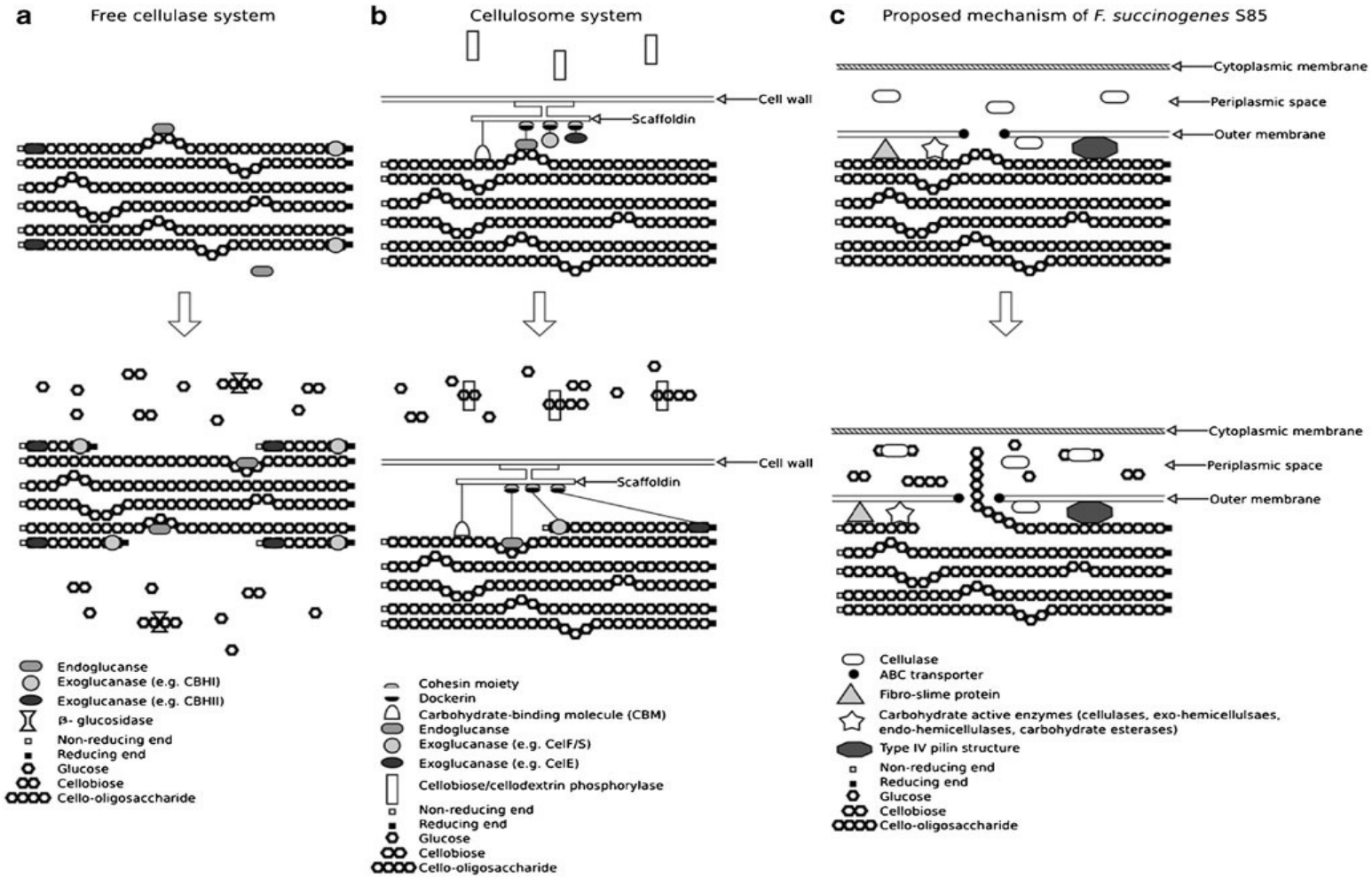


Figure 1.2. Mechanisms of microbial cellulose degradation. **a** Aerobic free cellulase system; **b** Anaerobic cellulosome system; **c** Proposed cellulolytic mechanism of *Fibrobacter succinogenes* (from Ransom-Jones *et al*., 2012).

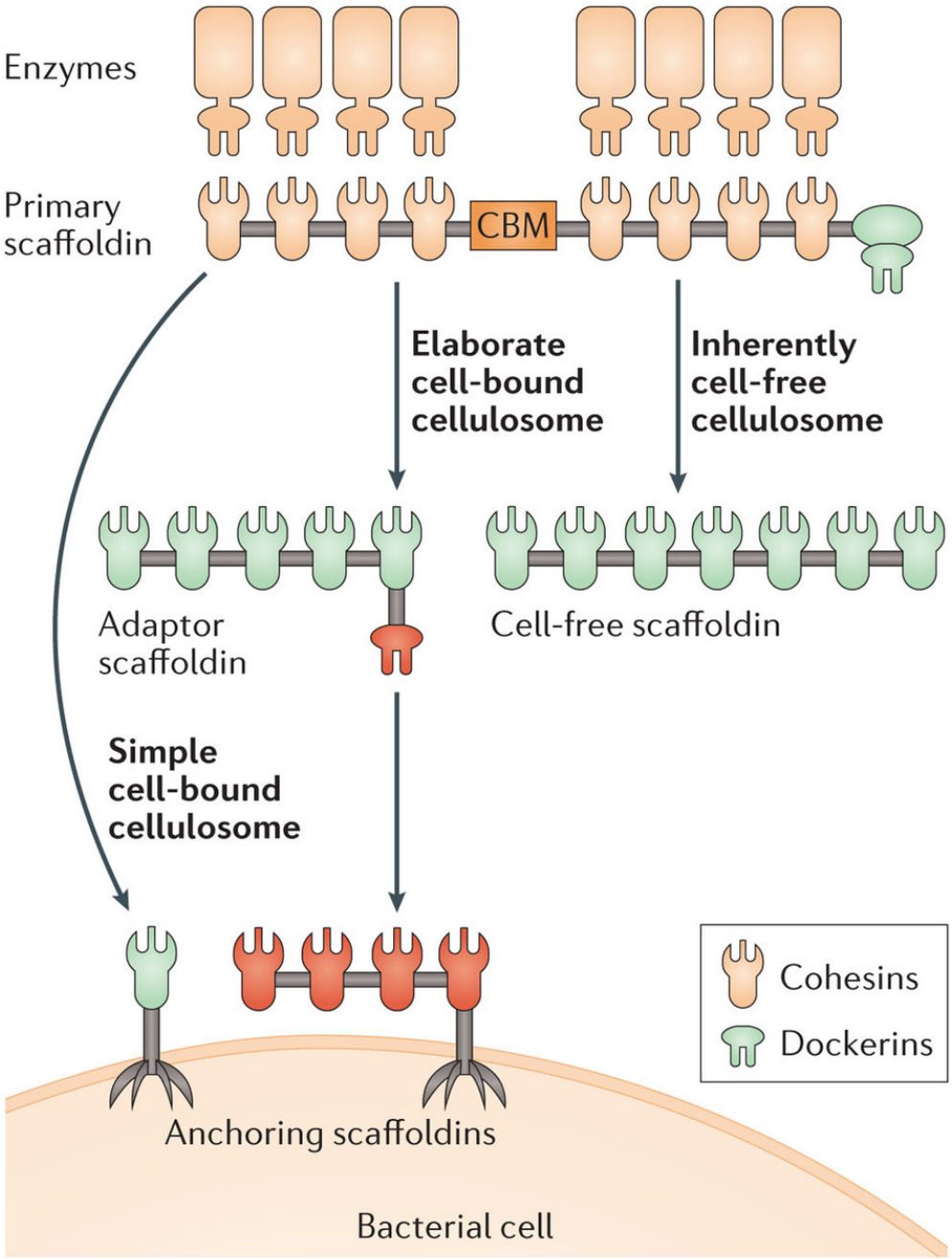


Figure 1.3. Structure and components of a cellulosome system. Primary scaffoldins can be bound directly to the cell through anchoring scaffoldins; primary scaffoldins can be bound to the cell through an intermediary adapter scaffoldin; or cellulosomes can exist in a free-state through attachment to a cell-free scaffoldin (from Artzi *et al*., 2017).

Unlike their aerobic counterparts, anaerobic cellulolytic microbes such as clostridia lack the ability to penetrate the substrate and it is believed that the evolution of cellulosomes offers them many advantages. Apart from causing rapid and efficient hydrolysis of cellulosic substrate, cellulosomes lead to the formation of contact corridors between the microbial cell wall and the substrate, leading to highly proficient uptake of cellulose degradation products with minimal loss of nutrients to the competing microbial community (Schwarz, 2001; Fontes & Gilbert, 2010). Presence of discrete catalytic subunits as well as the exact complement of catabolic subunits, along with catalytic modules required for the hydrolysis of interlinked polysaccharides such as hemicellulose facilitates optimal synergy for the decomposition of lignocellulosic biomass (Lynd *et al*., 2005; Artzi *et al*., 2017).

1.2.3.3 Proposed mechanism of *Fibrobacter succinogenes*

Recent evidence suggests that the anaerobic cellulose degrading bacterium *Fibrobacter succinogenes* uses a mechanism separate from the two reported so far (Fig. 1.2c) due to the apparent lack of genes encoding exocellulases and processive endocellulases (Wilson, 2008; Suen *et al.,* 2011). Emerging studies report that a protein complex, consisting of fibro-slime proteins and type IV pilin structures, present on the outer membrane mediates the attachment of the cell to the substrate. Individual cellulose chains are cleaved and transported into the periplasmic space via an ABC transporter, where endoglucanases hydrolyse the cellulose chains, thus eliminating the need for exocellulases and processive endoglucanases (Wilson, 2009). This novel insight is of industrial significance, as *F. succinogenes* demonstrates cellulolytic ability surpassing that of other rumen bacteria (Ransom-Jones *et al*., 2012). It is also interesting that genome analysis of the phylogenetically distinct aerobic soil bacterium

Cytophaga hutchinsonii suggests that it might use a mechanism similar to *F. succinogenes* (Wilson, 2009).

1.2.3.4 Glycoside hydrolases

Carbohydrate-Active enZymes (CAZymes) catabolise polysaccharides and oligosaccharides by cleaving the glycosidic bonds between carbohydrates or between a carbohydrate and a non-carbohydrate component, and can be broadly classified as Glycoside Hydrolases (GHs) or Polysaccharide Lyases (PLs) (Henrissat, 1991). GHs cleave glycosidic bonds through hydrolysis, while PLs use an elimination mechanism to cleave complex carbohydrates (Lombard *et al*., 2010). Both GHs and PLs are classified into families based on amino acid sequence similarities; as such a classification is necessary to reflect the structural features of the catalytic machinery and illustrate the evolutionary relationship between these enzymes, whilst also facilitating determination of molecular mechanistic information (Henrissat, 1991; Henrissat & Bairoch, 1993). Some families have been further grouped into 'clans', as the fold of the proteins can be better conserved than their amino acid sequence. The Carbohydrate-Active enZymes database (CAZy, Cantarel *et al*., 2009) is particularly useful, as it houses vast amount of information on catalytic and carbohydrate-binding modules associated with CAZymes, and it currently lists 133 GH families as well as 23 PL families.

GH families include enzyme classes displaying varying substrate specificities, and hence distinct catalytic functions; and one must be careful in making functional predictions based on sequence similarity alone. However, members of GH families display conserved features despite the differences in substrates hydrolysed, enabling the assignment of some GH families into broad substrate classes that facilitate

approximation of gene function (Henrissat & Davies, 1997; Cantarel *et al.*, 2012). For example, cellulases (EC 3.2.1.4) are represented in families GH5, GH6, GH7, GH8, GH9, GH10, GH12, GH26, GH44, GH45, GH48, GH51, GH74 and GH124; whereas xylanases (EC 3.2.1.8) are represented in families GH5, GH8, GH9, GH10, GH11, GH12, GH16, GH26, GH30, GH43, GH44, GH51 and GH62.

Due to the wide range of biological functions performed by carbohydrates, genes encoding CAZymes have undergone extensive convergent and divergent evolution, and as such their diversity reflects that of the substrate they catalyse. CAZymes usually exhibit a modular structure, where catalytic domains are supplemented with ancillary domains, including carbohydrate-binding modules, providing greater substrate specificity and enabling synergistic hydrolysis of complex carbohydrates (Cantarel *et al*., 2009).

Three classes of enzymes are utilised by cellulolytic microbes for complete hydrolysis of cellulose (Leschine, 1995; Lynd *et al*., 2002; Jalak *et al*., 2012):

- **Endoglucanases**, also known as 1,4-β-D-glucan-4-glucanohydrolases, are responsible for cleaving internal β-1,4 bonds at random within the amorphous regions. This disrupts the highly crystalline structure of cellulose fibres, generating oligosaccharides with reducing and non-reducing ends.
- **Exoglucanases** act sequentially from either the reducing or the non-reducing ends of the cellulose polymer, and include 1,4-β-D-glucan glucanohydrolases (also known as cellodextrinases) and 1,4-β-D-glucan cellobiohydrolases (also known as cellobiohydrolases). Glucose, by the action of glucanohydrolases, and cellobiose, by the action of cellobiohydrolases, are the major end products.
- **β-glucosidases**, also known as β-glucoside glucohydrolases, generate glucose and cellobiose moieties by the hydrolysis of cellodextrins.

Copper-dependent lytic polysaccharide monooxygenases (LPMOs) are widely considered as a major breakthrough in enzymatic cellulose decomposition (Vaaje-Kolstad *et al*., 2010; Quinlan *et al*., 2011) due to their powerful lytic activity. They were formerly classified as GH61 enzymes, but have now been placed under auxiliary activity (AA) enzyme families AA9, AA10, AA11 and AA13 (Levasseur *et al*., 2013). Although LPMOs are found both in bacterial and fungal kingdoms, they are most abundant in the genomes of saprotrophic fungi (Hemsworth *et al*., 2015), where AA9 LPMOs are exclusively found in fungi and act preferentially on cellulose (Beeson *et al*., 2015). Their active site co-ordinates a singular copper atom in a 'histidine brace', which comprises an N-terminal histidine residue along with an internal histidine and a tyrosine residue (Johansen, 2016). Their mechanistic action involves oxidative chain lysis of glycosidic bonds through acceptance of electrons from a variety of sources, including from plant-derived reducing agents such as diphenols. This renders the cellulosic substrate more susceptible to attack by other glycoside hydrolases, facilitating higher rate of hydrolysis through synergistic interactions (Johansen, 2016). To that effect, these enzymes demonstrate significant industrial importance, and are a major constituent of Novozymes' enzyme blend used for commercial consolidated bioprocessing (Leggio *et al*., 2015).

1.2.4 Anaerobic cellulose degradation

Anaerobic degradation of cellulosic biomass occurs in many terrestrial and aquatic ecosystems, including salt-water marshes, waterlogged soils, anaerobic sludges, wastewater treatment and municipal waste sites and marine, estuarine and freshwater sediments (Leschine, 1995). Moreover, microbial cellulose degradation in anoxic environments such as the rumen and herbivore gut, along with the intestines of

certain wood-feeding cockroaches and termites also relies on specialised composite microbial communities that have a significant role to play in the nutrition of the host (Slaytor, 1992; Warnecke *et al*., 2007; Brune, 2014). Due to the insoluble nature of cellulose, bacterial and fungal degradation occurs extracellularly. This results in cellulose hydrolysis products being utilised by neighbouring microbes, facilitating a broad range of microbial interactions within the environment (Leschine, 1995; Barlaz, 1997).

1.2.4.1 The rumen

Since herbivores are unable to degrade the large amounts of plant polysaccharides they ingest as a part of their diet (Bauchop, 1979; Dehority, 2003), they have evolved symbiotic relationships with microbes that harbour the necessary enzymes for lignocellulosic biomass degradation. These symbiotic bacteria, fungi and protozoa provide the host with easy to assimilate forms of carbon and energy, in return for a constant source of plant matter along with optimal conditions for their survival and proliferation (Leschine, 1995). Due to the significant role cellulolytic microbes play in host nutrition, microbial ecology of cellulose hydrolysis in rumen and herbivore gut is well understood, and much of our current knowledge of anoxic cellulose decomposition stems from here (Tajima *et al*., 1999; Tajima *et al*., 2000; Tajima *et al*., 2001; Michalet-Doreau *et al*., 2001; Krause *et al*., 2003).

The temperature of the rumen is maintained at 39ºC and it harbours a dense microbial population, estimated at ~10^{9}-10^{10} bacteria ml^{-1}, ~10^{5}-10^{6} protozoa ml^{-1} and ~10^{1} fungal zoospores ml^{-1} (Theodorou *et al*., 1990). As such, the rumen behaves like a continuous culture system, where fermentation of the plant biomass leads to the formation of volatile fatty acids such as acetate, propionate and butyrate. Despite the

entry of small amounts of air during feeding, strict anoxia is maintained in the rumen as facultative anaerobic bacteria swiftly utilise the oxygen, keeping the redox potential of the rumen between -250 and -400 mV (Hobson & Wallace, 1982).

Although over 200 species of rumen bacteria have been reported, culture-based techniques utilised for the isolation of cellulolytic bacteria suggested *Fibrobacter succinogenes*, *Ruminococcus albus* and *Ruminococcus flavefaciens* to be the predominant species (Hungate, 1966). Molecular techniques reliant on amplification of the 16S rRNA gene of cellulose degrading rumen bacteria have generally supported this observation (Tajima *et al*., 2001; Denman & McSweeney, 2006; Shinkai & Kobayashi, 2007), with *Ruminococcus* accounting for 8% and *Fibrobacter* accounting for 1% of the bacterial abundance (Stevenson & Weimer, 2007). In addition, anaerobic cellulolytic fungi have been reported to contribute heavily towards lignocellulosic biomass hydrolysis, whereas out of the over 100 species of rumen protozoa have been identified, only a few have been documented to have any cellulolytic function as most protozoa have been hypothesised to grow by predation on rumen bacteria (Russell *et al*., 2009). The highly complex community of microbes inside the rumen includes bacterial species that rely on the generation of fermentation products (lactate and succinate) from other microbes as well as methanogens that utilise formate, H_2 and CO_2 for the generation of methane gas (Trinci *et al*., 1994; Leschine, 1995).

Anaerobic microbial communities that degrade lignocellulose are structurally and metabolically complex, and studies addressing the abundance and relative activity of their constituent microbial groups have mainly focussed on the herbivore gut (Leschine, 1995; Rubin, 2008). As a consequence, relatively unexplored anoxic environments such as landfill sites can be viewed as a potential reservoir of novel cellulolytic agents, as difficult to isolate obligate anaerobes may be present.

1.2.4.2 Landfill sites

Landfill sites have commonly been used for the disposal of wastewater treatment sludge, agricultural and industrial waste as well as Municipal Solid Waste (MSW) (Kjeldsen *et al*., 2002), which includes kitchen and garden waste, paper, card and plastic packaging, and miscellaneous combustible and non-combustible refuse from residential and industrial sources (USEPA, 1994). Modern day landfill sites comprise highly structured and regulated facilities designed in order to minimise their environmental impact by controlling the flow of leachate and the gas generated (Barlaz, 1997). Leachate collection in a landfill site is facilitated by the construction of a liner system at the base of a site, consisting of thick layers of clay, sand and polyethylene. This set-up serves to allow leachate collection whilst also preventing contamination of groundwater (Daniel, 1993).

Landfill sites offer a cheaper disposal option to composting and incineration, whilst also ensuring that the microbial degradation of waste occurs in a controlled manner within a contained environment (Renou *et al*., 2008). Moreover, methane gas produced from the operation of landfill sites has been utilised for the generation of electricity (Jaramillo & Matthews, 2005; Bove & Lunghi, 2006), as it is possible to operate and decommission landfill sites sustainably relying solely on the generated gas as a source of energy (Barlaz *et al*., 2003a,b). Of the ~163 million tonnes of waste disposed of in the UK in 2006, over 70% of it ultimately made its way to the landfill, either directly (~69 million tonnes) or indirectly (~47 million tonnes) (Environment Agency, 2006). Just as in the UK, global reliance on landfill sites is expected to decrease due to recent efforts in waste management. However, only a proportion of MSW can be recycled appropriately and as such, landfill sites will continue to be an

integral part of MSW disposal (Barlaz, 2006). The composition of all municipal waste originating in England during 2006-2007 has been presented in Table 1.2.

Even though the composition of waste in a landfill site can be immensely heterogeneous (Burnley, 2007), Barlaz (2006) estimated that lignocellulose constituted the major proportion of organic matter in the residential refuse entering MSW sites in the USA (~ 57% dry weight). Table 1.3 presents data on the estimated organic composition of certain household waste subcomponents in the USA.

Table 1.2. Estimated composition of all municipal waste in England during 2006-2007 (DEFRA, 2009).

Component	**% estimated composition (by weight)**
Food waste	17.8
Garden waste	14.1
Other organic	1.7
Paper and card	22.7
Glass	6.6
Metals	4.3
Plastics	10.0
Textiles	2.8
Wood, furniture and mattresses	5.3
Waste electrical and electronic equipment	2.2
Sanitary and hazardous	3.0
Miscellaneous combustible	2.4
Miscellaneous non-combustible	2.8

Table 1.3. Estimated organic composition of certain household waste subcomponents in the USA (Barlaz, 1996).

Component	**Estimated composition (% of dry weight)**			
	Cellulose	Hemicellulose	Lignin	Volatile solids
Garden waste	25.7	13.0	34.9	90.6
Food waste	50.8	6.7	9.9	92.0
Office paper	87.4	8.4	2.3	98.6
Coated paper and newsprint	45.4	8.7	19.5	86.4
Corrugated boxes	57.3	9.9	20.8	98.2
Mixed refuse	42.5	10.5	17.6	76.9

1.2.4.2.1 Microbiology of landfill sites

Landfill sites harbour complex microbial communities due to optimal conditions created by the presence of effluent organic and inorganic nutrients, surfaces for adherence and colonisation, neutral pH, adequate moisture and slightly raised temperature (Palmisano and Barlaz, 1996). The formation of microenvironments due to the presence of suspended matter within the leachate leads to multiple distinct microbial metabolic processes occurring simultaneously within the highly diverse interacting communities (Palmisano and Barlaz, 1996). It has been proposed that such communities of interdependent anaerobic microbes mediate the metabolism of lignocellulosic biomass into methane, similar to the mechanism of its decomposition observed in other anoxic environments, including the rumen, wastewater sludge digesters and salt water marshes (Barlaz, 1996).

An overview of the process has been presented in Figure 1.4. Briefly, the decomposition is initiated by the hydrolysis of complex substrates such as polysaccharides and proteins into simpler compounds such as sugars, amino acids, long-chain carboxylic acids and glycerol. Fermentation of hydrolysis products generates alcohols, short-chain fatty acids (acetate, propionate and butyrate), H_2 and CO_2, followed by acetogen-mediated oxidation of the fermentation products into acetate, H_2 and CO_2 (Barlaz, 1996). The oxidation reaction is achieved in syntrophy with the scavenging of H_2 by methanogens and Sulphate Reducing Bacteria (SRB), as the reaction is thermodynamically unfavourable at high concentrations of H_2. Anaerobic degradation is concluded as methanogens ultimately generate methane gas from acetate, H_2 and CO_2 (Barlaz *et al*., 1989).

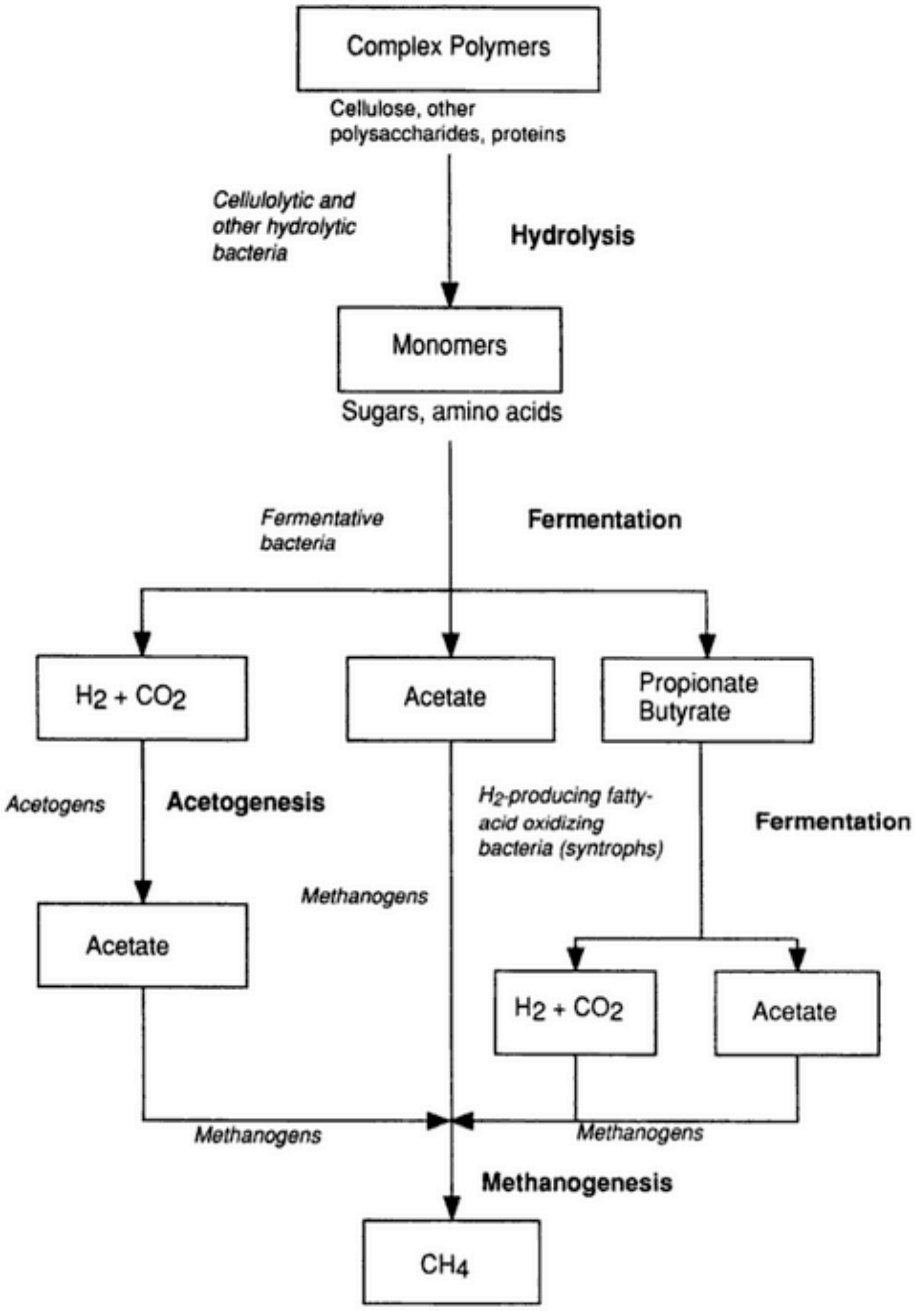


Figure 1.4. Process of the bioconversion of complex organic matter into methane and carbon dioxide by the anaerobic microbial community in landfill leachate (from Brock *et al*., 1994).

Much of our knowledge relating to the degradation of waste in landfill sites has been derived from simulations performed in laboratories under optimum conditions, which are governed by abiotic factors including temperature, pH, moisture levels and waste composition. Since Farquhar and Rovers (1973) first proposed the mechanism of anaerobic degradation of MSW in landfill sites, Barlaz *et al*. (1989) and Kjeldsen *et al*. (2002) have further detailed the pathway by performing laboratory simulations. Such simulations were necessary as *in vivo* MSW decomposition in landfill is a relatively slow process and relies on the formulation of optimum growth conditions for the establishment of a microbial population, which are variable due to the heterogeneous nature of the deposited waste. The process of landfill waste decomposition has been detailed in (Fig. 1.5), and the four phases involved are (Barlaz *et al*., 1989; Barlaz, 1996):

- **Aerobic phase**
- **Anaerobic acid phase**
- **Accelerated methane production phase**
- **Decelerated methane production phase**

Microbial ecology studies conducted in our lab have detected novel cellulose-degrading relatives of the bacterial genus *Fibrobacter* and the fungal genus *Neocallimastix* that form an environmentally significant proportion of the microbial community in landfill leachate, freshwater and marine sediments, species of which were previously thought to be obligate gut inhabitants (Van Dyke & McCarthy, 2002; Lockhart *et al*., 2006; de Menezes *et al*., 2008; McDonald *et al*., 2008; McDonald *et al*., 2009; McDonald *et al*., 2012).

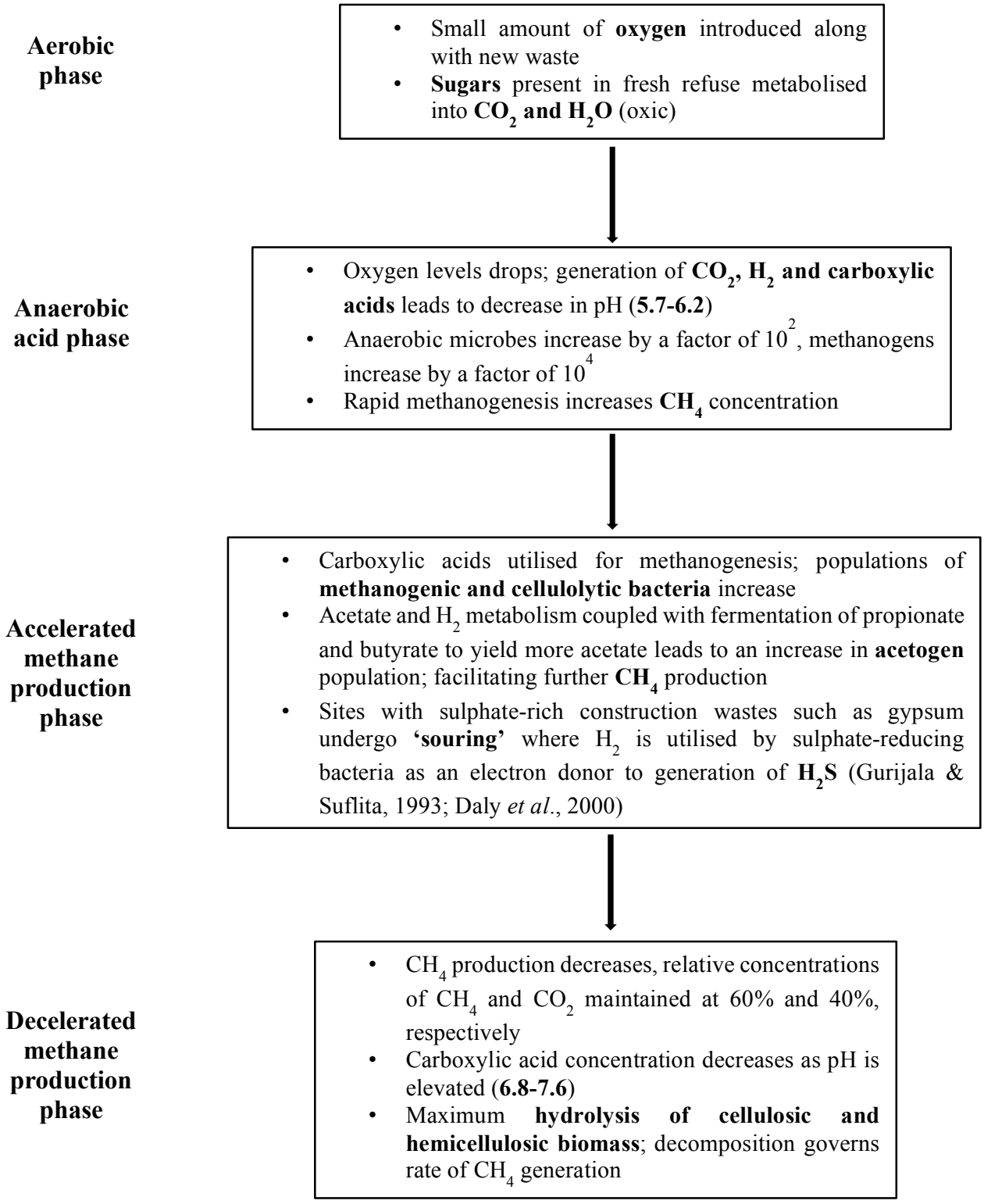


Figure 1.5. The four phases involved in the anoxic decomposition of landfill waste (Barlaz *et al*., 1989; Barlaz, 1996).

1.2.5 Cellulolytic microbes

Microorganisms that display the ability to degrade cellulose are found distributed across all domains of the universal tree of life (reviewed by Lynd *et al*., 2002). Anaerobic fungal group *Chytridiomycetes* are known degraders of cellulose in the herbivore gut (Orpin, 1975), whereas aerobic fungal groups including *Ascomycetes*, *Basidiomycetes* and *Deuteromycetes* consist of a large number of cellulolytic species (Lynd *et al*., 2005). Ciliated protozoa, including *Nyctotherus* and *Trichonympha*, make up a significant proportion of the cellulose-degrading microbial population in the intestinal tract of certain termites and wood-feeding cockroaches (Gijzen *et al*., 1994; Lynn, 2010). Presence of considerable diversity means that cellulolytic bacteria can be divided into three distinct physiological groups; (1) fermentative anaerobes, generally consisting of Gram-positive bacteria such as *Ruminococcus* and *Clostridium*, but also include phylogenetically related (*Acetivibrio* and *Butyvibrio*) and unrelated (*Fibrobacter*) Gram-negative species; (2) Gram-positive aerobes, including *Cellulomonas* and *Thermobifida*; and (3) aerobic gliding bacteria, including *Cytophaga* and *Sporocytophaga* (Lynd *et al*., 2002). Identification of cellulolytic enzymes in *Archaea* has also been reported (Ando *et al*., 2002; Birbir *et al*., 2007).

1.2.5.1 Anaerobic cellulolytic fungi

Anaerobic cellulolytic fungi exhibit exceptional ability to hydrolyse lignocellulosic biomass, as they possess a broad range of cellulases, hemicellulases and esterases, which are produced in copious amounts (Leschine, 1995). They are widely regarded as one of the most potent cellulose degraders in the known biological world, with the cellulases synthesised by *Neocallimastix frontalis* known to demonstrate greater specific activity than any other cellulase observed to date (Wilson & Wood,

1992a; Wilson & Wood, 1992b). Their ability to gain access to lignocellulosic biomass by penetrating plant tissue such as the cuticle allows them to perform rapid and complete hydrolysis of the substrate.

Surprisingly little is known about the unique behaviour and life cycle of these fungi. It is thought that they reproduce asexually by medium of motile zoospores that migrate to cellulosic surfaces in order to germinate. This results in the development of a thallus containing either monocentric or polycentric sporangia, leading to the formation of a highly branched rhizomycelium that facilitates extensive plant cell wall degradation (Haitjema *et al*., 2014). Monocentric fungi propagate themselves only by the release of zoospores, whereas polycentric fungi can do so by rhizoidal fragmentation as well (Theodorou *et al*., 1996). Davies *et al*. (1993) reported the recovery of viable cultures from herbivore faeces, which was unexpected as these microbes were thought to be obligate anaerobes. A possible explanation is that fungal migration between anoxic environments might be facilitated by a third life cycle stage comprising of aero-tolerant spore-like structure, an observation supported by Brookman *et al*. (2000b).

Most studies relating to anaerobic fungi have focussed on the herbivore gut, where they have been reported to represent around 8% of the biomass (Kemp *et al*., 1984), and yet have been estimated to account for an astonishing 40-70% of lignocellulose degradation (Akin & Rigsby, 1987). They are reported to be the primary colonisers of ingested plant biomass (Theodorou *et al*., 2005), which is remarkable considering that bacteria and protozoa in the rumen can outnumber these microbes by orders of magnitude. It is then, perhaps, not surprising that cellulolytic bacteria were considered to be the primary plant biomass degraders in the herbivore rumen before the discovery of these fungi (Borneman & Akin, 1994).

It was long believed that all fungi required oxygen for survival, something that is likely to have delayed the identification of anaerobic fungi from rumen (Trinci *et al*., 1994) as microbiologists historically proceeded to strain the rumen fluid before isolating microbes, discarding plant matter and the fibre-associated fungal biomass (Bauchop, 1981). They have since been isolated from over 50 large mammalian herbivores including cow, sheep, horse, buffalo, yak, camel, kangaroo, llama, rhinoceros and elephant (Ljungdahl, 2008); while DNA has been identified in others including zebra, donkey, giraffe, antelope, gazelle, deer and elk, as well as the reptilian herbivore iguana (Liggenstoffer *et al*., 2010).

Traditionally, genus level classification of anaerobic fungi was performed on the basis of cellular morphology, whereas species level classification was performed using electron microscopy to determine the shape and location of organelles (Ho & Barr, 1995). Billon-Grand *et al*. (1991) reported that anaerobic fungi are a homogenous group characterised by their extremely low GC content (13-22%), based on the analysis of 18S rRNA and enzyme-coding genes. However, 18S rRNA genes tend to be highly conserved, which led to Brookman *et al*. (2000a) to propose the use of the more variable and yet suitably conserved internal transcribed spacer 1 (ITS1) region for the purpose of the classification of anaerobic fungi. A new isolate can be typically identified by a sequence divergence or approximately 3% in the ITS1 region sequence (Nilsson *et al*., 2008). This has led to the classification of 6 genera of anaerobic fungi, viz. *Neocallimastix*, *Caecomyces*, *Piromyces*, *Anaeromyces*, *Orpinomyces* and *Cyllamyces* (Fig. 1.6).

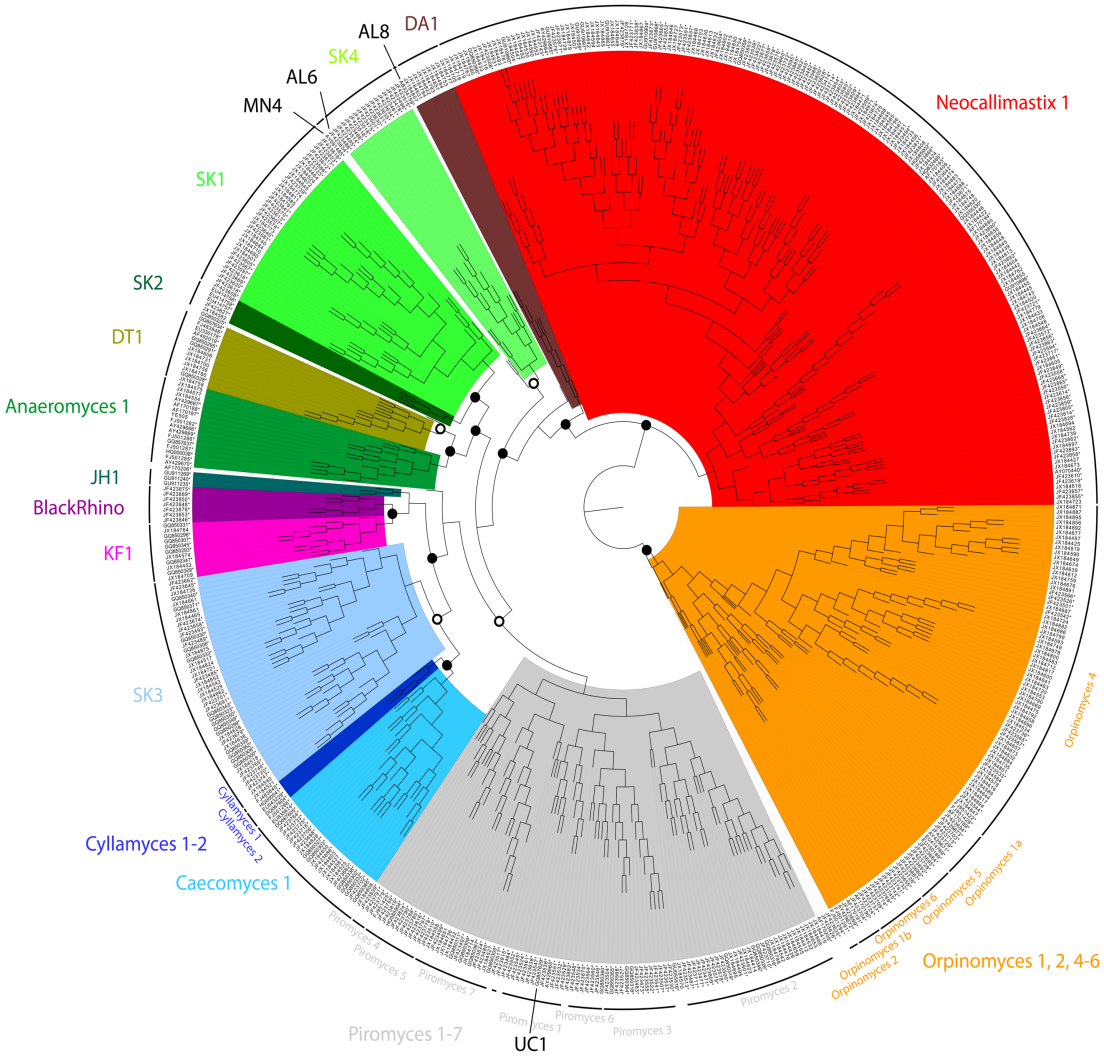


Figure 1.6. An illustration of anaerobic fungal phylogeny. Profile Neighbour Joining tree of anaerobic fungal Phylum Neocallimastigomycota generated using sequence and structure data from 1120 complete ITS1 sequences with 1,000 bootstrap replicates. The six known genera can be seen (*Neocallimastix*, *Caecomyces*, *Piromyces*, *Anaeromyces*, *Orpinomyces* and *Cyllamyces*). Open circles indicate bootstrap values 50-90, closed circles indicate bootstrap value >90 (from Koetschan *et al*., 2014).

Genome analysis of an isolate of the genus *Orpinomyces* has recently been reported using a combination of Illumina and PacBio SMRT technologies. As expected, the genome was determined to be large (~101 Mb, comprising of over 16,000 genes) with an extremely low GC content of only 17%. Analysis suggested that multiple instances of horizontal gene transfer from bacteria are responsible for the acquisition of extremely potent lignocellulolytic machinery, facilitated by their separate evolutionary trajectory in the rumen. The strain C1A was also noted to be able to perform simultaneous saccharification and fermentation of cellulosic and hemicellulosic biomass, with its ability significantly enhanced by mild pre-treatment of the substrate (Youssef *et al*., 2013).

To date, *Neocallimastigales* fungi have not been isolated from any environment other than the herbivore gut, and it was assumed that the presence of these microbes was restricted to such an environment (Haitjema *et al*., 2014). There is however molecular evidence for the presence of these fungi in landfill sites through the targeting of 18S rRNA gene (Lockhart *et al*., 2006; McDonald *et al*., 2012), where the occurrence of these fungi correlated with the cellulose content of the landfill. This suggests a wider ecological distribution of these fungi in anoxic environments and a potentially crucial role in cellulose hydrolysis in general. As such, it is important to isolate these anaerobic fungi from other anoxic environments to better elucidate their physiology, and to be able to perform whole genome sequencing and transcriptomic analysis.

1.2.5.2 Anaerobic ciliated protozoa

Ciliated protozoa, belonging to the phylum Ciliophora, vary greatly in their appearance, displaying a size range between 10 μm to 4,500 μm and a variety of unusual shapes (Lynn & Corliss, 1991). Both free-living and symbiotic forms of ciliates

have been described globally from a variety of habitats, including oceans (Lynn & Montagnes, 1991), freshwater lakes (Taylor & Heynen, 1987), hypersaline lagoons (Garcia & Neill, 1993) and terrestrial soils (Buitkamp, 1977). The life cycle of a typical ciliate consists of an asexual stage for growth and division, a sexual stage for exchange of genetic material, and a cryptobiotic stage facilitating stress-induced cyst formation (Lynn & Corliss, 1991). All ciliates are heterotrophic (Finlay & Fenchel, 1996), and are briefly characterised by three main features:

- The presence of cilia over the body surface, variable in number and arrangement, derived from kinetosomes
- Nuclear dimorphism, where the macronucleus controls the physiological and biochemical functions of the cell, and the micronucleus acts as the germ-line reserve
- Utilisation of conjugation as a sexual process, where temporary fusion of partners leads to the exchange of genetic material

Ciliates are believed to have evolved from flagellates (Lynn & Small, 1981), and their current taxonomic classification divides them into 2 subphyla, Postciliodermatophora and Intramacronucleata, consisting of 2 and 9 classes, respectively (Fig. 1.7). Such a classification was strongly supported by SSU rRNA-based molecular phylogenetic studies as well as morphological characteristics, as it has been suggested that a comprehensive classification of ciliates should ideally involve molecular evidence, morphological patterns and cortical ultrastructure (Agatha *et al*., 2005). Ciliates displaying somatic kinetids with postciliodesmata were placed in the former subphylum, while those demonstrating macronuclear division by intramacronuclear microtubules were placed in the latter (Lynn, 1996a). The ability to degrade cellulose is observed only in a subset of endosymbiotic protozoa: certain

ciliates belonging to class Litostomatea and class Armophorea. Degradation of fibrous feed occurs intracellularly in these protozoa following ingestion of the lignocellulosic biomass, as opposed to the cellulosome model observed in anaerobic bacteria and fungi.

Ciliates in the class Litostomatea are divided into 2 subclasses: Haptoria and Trichostomatia. While haptorians are free-living predatory ciliates, the trichostomes are found as endosymbionts in a variety of metazoans, ranging from reptiles to mammals where they are associated with phagocytosis of bacteria and plant matter. Some or all isotrichids, paraisotrichids, ophryoscolecids, entodiniomorphids, cycloposthiids and buetschliids have been reported from cattle (Gocmen *et al*., 2001), yak (Guirong *et al*., 2000), deer (Dehority, 1995), horses (Bonhomme-Florentin, 1994), South American capybara (Ito & Imai, 2000), elephants (Timoshenko & Imai, 1997) and Australian macropodid marsupials (Cameron & O'Donoghue, 2003b), amongst others. Further to their global distribution, trichostome abundance has been estimated to range from 10^4 ml^{-1} in yak to around 10^6 ml^{-1} in the collared peccary (Carl & Brown, 1983). Important endosymbiotic ciliates include *Entodinium*, *Isotricha*, *Polyplastron*, *Epidinium*, *Ophryoscolex* and *Eudiplodinium*.

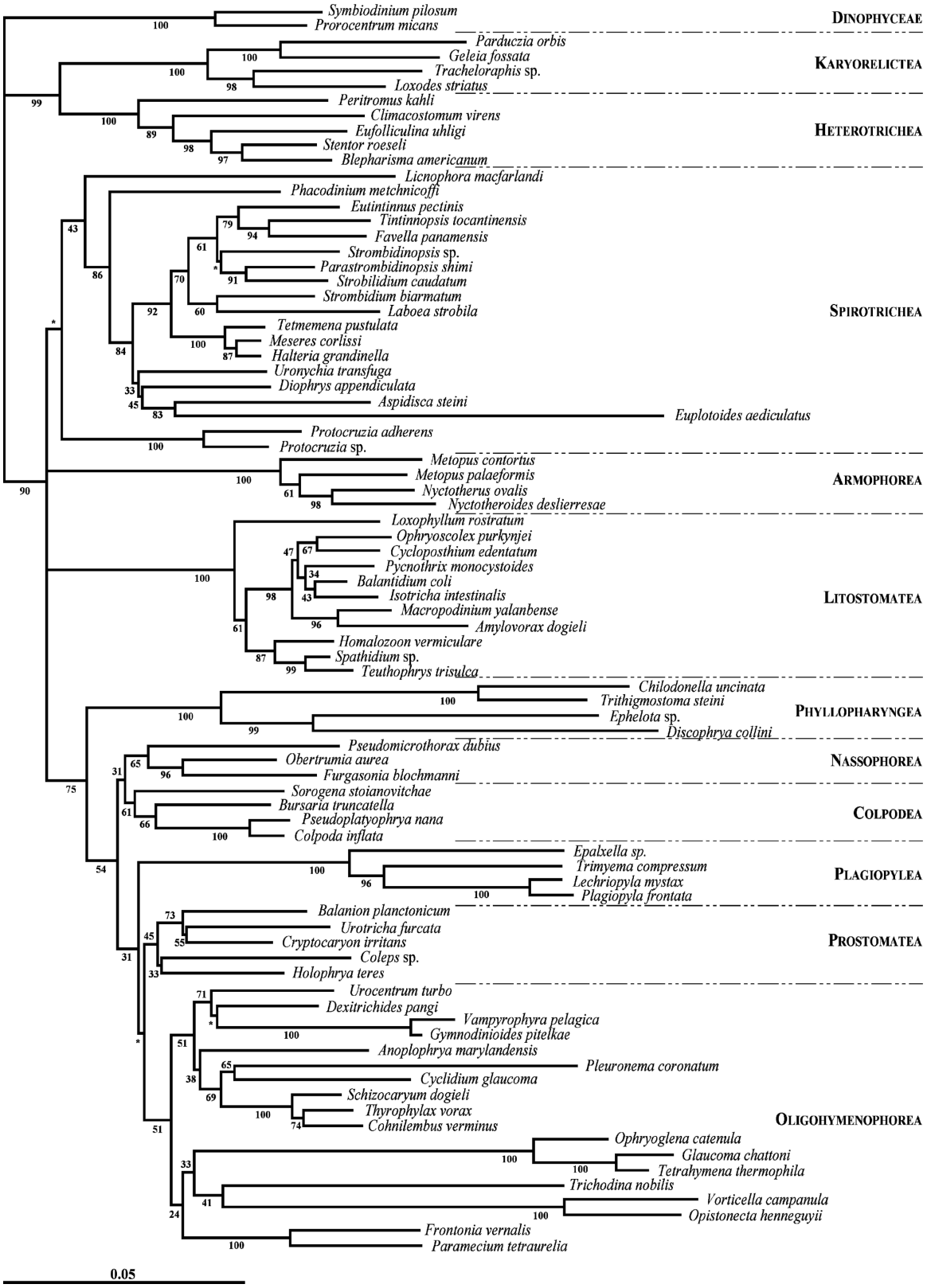


Figure 1.7. A phylogenetic tree illustrating the molecular phylogeny of anaerobic ciliate Phylum Ciliophora based on SSU rRNA gene sequences (From Lynn, 2010).

The role of rumen protozoa in the degradation of lignocellulosic biomass is still unclear, despite Hungate (1942) first demonstrating the cellulose degrading capability of certain large entodiniomorphid ciliates isolated from the rumen. Classical microbiology studies supported such observations, as Yoder *et al*. (1966) reported an increased rate of cellulose degradation *in vitro* in the presence of protozoa, whereas Luther *et al*. (1966) and Veira *et al*. (1983) observed that cellulose hydrolysis was 10-20% lower in defaunated lambs. Demeyer (1981) suggested that ciliated protozoa could be responsible for as much as 30-50% of total cellulose degradation in sheep. In contrast, some studies reported no significant effect (Bonhomme, 1990; Jouany, 1994) or even lower digestibility (Eadie & Gill, 1971) of fibrous lignocellulose in faunated sheep compared with ones that had been defaunated. It was long believed that the cellulose-degrading capability associated with protozoa in the rumen was due to the bacteria engulfed by them, as rumen ciliates are important predators of bacteria. However, further studies on expressed cellulases and xylanases later concluded that these enzymes were in fact of protozoan origin (Devillard *et al*., 1999; Michalowski *et al*., 2003; Stan *et al*., 2006).

Ciliates belonging to the class Armophorea characteristically have their mitochondria transformed into hydrogenosomes, as these organelles serve to provide hydrogen to the endosymbiotic bacteria, predominantly methanogenic, associated with them. Although their distribution is global, these ciliates demonstrate limited habitat diversity due to their anaerobic metabolism. Their habitat is restricted to anoxic aquatic (Fenchel, 1993) and terrestrial sediments (Foissner, 1987), as well as digestive tracts of certain invertebrates (Hackstein & Stumm, 1994), reptiles and amphibians (Affa'a *et al*., 1995). Finlay and Fenchel (1991) report the presence of free-living armophorids in a variety of MSW landfill sites in the UK, where they undergo encystment in response

to starvation and water loss. Ciliates in order Clevelandellida are obligate endosymbionts found commensal to be in a wide range of hosts: *Nyctotherus* is found in myriapods and insects such as termites (Hackstein & Stumm, 1994), *Nyctotheroides* is found in frogs and toads, and *Clevelandella* is found in wood-feeding cockroaches and termites (Gijzen & Barugahare, 1992). Endosymbiotic armophorids have been reported to contribute towards a significant rise in the growth rate and body weight of their invertebrate hosts (Gijzen & Barugahare, 1992). Moreover, Gijzen *et al*. (1994) established that *Nyctotherus ovalis* was responsible for the major cellulolytic and methanogenic activity in the hindgut of cockroach *Periplaneta americana*, as a close correlation was observed between protozoal numbers and the cellulase activity, as well as the methane production, of the insect.

While some literature exists outlining the distribution of armophorid ciliates in MSW landfill sites, the research has primarily focused on the methanogenic endosymbionts of these anaerobic ciliates. Consequently, little is known about any potential role the armophorids might have in cellulose decomposition in landfill sites rich in lignocellulosic biomass.

1.3 Sampling sites and experimental design

Distinct anoxic environments were chosen as the subject of this study, as cellulose is known to be available as a carbon source and is broken down by anaerobic microbial communities present within them. A large number of previous studies have focussed on particular environments and niches where cellulolytic microbes have been known to be present in high numbers, leading to the identification of interesting cellulases and GHs. Most metagenomic studies have targeted the rumen, particularly focussing on the microbes associated with lignocellulosic plant biomass in the bovine

rumen, and this has resulted in the isolation of numerous carbohydrate active enzymes (e.g. Brulc *et al.*, 2009; Hess *et al*., 2011). Cellulose baits were used in this study as a means of enrichment as previous studies have resulted in the detection of interesting microbes and genes (Edwards *et al*., 2010; de Menezes *et al*., 2012). This approach allows for biofilm formation by genuinely cellulolytic microbes (Fig. 1.8), and as such might improve the chances for isolation of novel cellulolytic genes from environments where cellulose-degrading microbes exist as part of highly diverse communities.

One of the environments studied is the anoxic leachate associated with MSW, collected from Bromborough Dock and Bidston Moss landfill sites (Wirral, UK) in this case. Both the sites were no longer in use and had been capped, and as such, a system of 'risers' allowed sampling by pumping the leachate up to the surface. Leachate is typically associated with dissolved and suspended waste matter, with cellulose and hemicellulose estimated to make up to 50-60% of the total biodegradable biomass in a landfill site (Barlaz, 2006). Landfill leachate previously collected from Bromborough Dock has been demonstrated to harbour a significant cellulolytic microbial community. Molecular ecology studies have also detected members of the cellulolytic *Fibrobacter* spp. that are distinct, yet closely related, to those resident in the rumen (McDonald *et al*., 2008; McDonald *et al*., 2009).

The freshwater lake Esthwaite Water (Lake District National Park, UK) was selected for a functional metagenomic survey of its microbial community, since it is situated in a fertile valley and has been documented to be one of the more productive lakes in the area (George *et al*., 2000). Sampling was performed at the deepest part of the lake, which is approximately 14 m deep (subject to seasonal variation), where the sediment and the lower part of the water column are anoxic in nature. McDonald *et al*. (2009) and de Menezes *et al*. (2008; 2012) have previously undertaken work associated

with the molecular ecology of cellulose degradation in Esthwaite Water, focussing their investigation on *Fibrobacter* spp. and *Micromonospora* spp., respectively. Dr James Houghton followed up those studies with metagenomic and metatranscriptomic analyses of cellulose degradation in lake sediment, and a similar 'omics'-based survey of the cellulolytic community in landfill leachate is presented in this thesis.

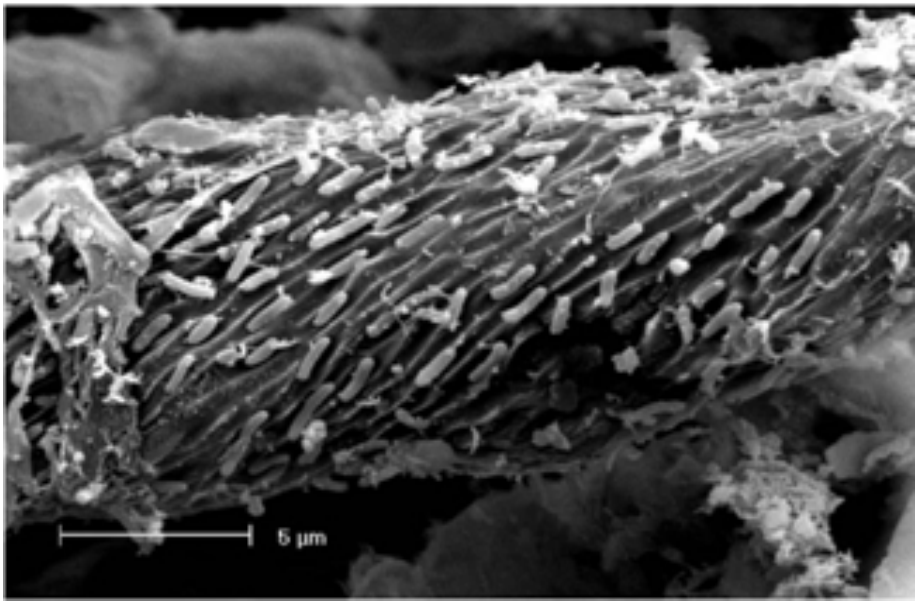


Figure 1.8. Scanning electron micrograph of rod-shaped bacteria colonising the surface of a cellulose bait *in situ* in the Irish Sea (from Edwards *et al*., 2010).

1.4 Aims of the project

It was hypothesized that understudied anoxic environments could harbor potentially novel cellulolytic agents that display significant hydrolytic capability, with a view to potential application in industry. The overall aim of the project is to characterise the structure and function of the indigenous microbial population responsible for anaerobic cellulose degradation in landfill. This will be achieved by elucidating the phylogenetic constitution of the microbial community, as well as mining for genes expressed in this environment that are involved in cellulose degradation.

The project has the following specific aims:

- To attempt to culture anaerobic fungi belonging to the order *Neocallimastigales* from landfill leachate, as previous data (Lockhart *et al*., 2006) points towards the presence of these exceptional cellulose degraders in landfill sites.
- To perform high throughput sequencing of the metagenome (total DNA) and the metatranscriptome (total mRNA) from the microbial community that colonises and degrades cellulose baits incubated in landfill leachate, with a view to deciphering the phylogenetic structure and the enzyme-coding potential of the population. A comparison between the metagenome and metatranscriptome will enable us to establish which groups of microbes are actively involved in crystalline cellulose colonisation and potentially decomposition.
- To determine the abundance and the diversity of expressed glycoside hydrolases in the landfill leachate community, achievable through high throughput sequencing of the community metatranscriptome.
- To produce a contemporaneous metagenomic library using high molecular weight DNA extracted from the colonised cotton using a fosmid vector system, with a view to cloning whole functional genes and gene clusters. Functional

screening of fosmid libraries produced from colonised cotton incubated in anoxic landfill leachate, as well as in anoxic lake sediment, would allow for isolation, overexpression and characterisation of any successfully cloned glycoside hydrolases.

- To develop a suitable method for extraction of good quality, downstream application-ready RNA from microbial communities resident in anoxic, hyperalkaline environments.

Chapter 2

Methods

2.1 Environmental sampling

Samples were collected from anoxic environments for the purpose of extracting nucleic acids and attempting to isolate anaerobic microorganisms, and included landfill leachates and sediment from a hyperalkaline lagoon. Both the landfill leachates and the hyperalkaline lagoon sediment were used directly, and as matrices for incubation of cellulose baits.

2.1.1 Generation of crystalline cellulose baits

Crystalline cellulose baits were generated using a method adapted from Wood (1988) for simplification. The method described by Wood is necessary for the preparation of extremely pure cotton for quantitative chemical assays, whereas this study only required the generation of crystalline cellulosic substrate to facilitate colonisation by microbes and act as a method of enrichment. Cotton string was dewaxed in a soxhlet apparatus by successive boiling in chloroform, 95% ethanol and twice in distilled water for eighteen hours each. The string was then soaked in cold distilled water overnight and then allowed to dry. The dewaxed cotton was devoid of impurities such as naturally associated waxes, and consisted of 99% pure crystalline cellulose. The cotton was placed into nylon mesh bags (Figure 2.1) before being used for *in situ* environmental sampling. The inert nylon bags were chosen to hold the cotton in place,

whilst allowing for the establishment of an enriched microbial community during incubation.

2.1.2 Landfill leachate sampling

Landfill leachate was collected from Bromborough Dock and Bidston Moss municipal waste landfill sites in the Wirral (Northwest England), and was stored in 10 L carboys (Nalgene) at room temperature in the laboratory (Figures 2.1). While waste tipping has ceased and both landfill sites have been capped with top soil, they remained active for methane generation. The Bromborough Dock leachate samples comprised a combination of leachate from 'risers' 3 and 4 (Figure 2.2), as it was impossible to sample those risers independently due to internal issues with the pumping system. Triplicate samples were acquired from this site, and are referred to as BD1, BD2 and BD3. The Bidston Moss site does not have a pumping system coupled with risers. As such, leachate samples were acquired manually from five different boreholes, and are referred to as BM3E, BM3F, BM3G, BM3H and BM1J. Each carboy was filled to the brim with leachate to ensure that anoxic conditions were maintained, and the cotton baits were incubated within the carboys for a period of 3-12 months, and harvested at intervals for nucleic acid extraction.

2.1.3 Hyperalkaline sediment sampling

pH 12.0-13.0 sediment was collected from a pond in a valley in Harpur Hill (Buxton, English Peak district) which was situated downstream from an old lime kiln and displayed heavy calcium deposition due to the extensive leaching of $CaOH_2$ and $CaCO_3$ (Figure 2.2.). The sediment was obtained from the soil-water interface where a large amount of plant matter was found to be dead and decaying, and was sampled from 9-

12 inches below the ground. Additionally, dewaxed cotton baits in nylon bags were also incubated in the sediment 18-24 inches below the ground for a period of 3 months before being retrieved for nucleic acid extraction following the potential establishment of an enriched microbial community. Both the sediment and the colonised cotton were collected in 50 ml centrifuge tubes (Greiner) that were filled to the top either with sediment alone or with the cotton supplemented with sediment, to ensure the maintenance of anoxic conditions. The centrifuge tubes were then placed inside an anaerobic container for the duration of the transportation back to the laboratory, before being processed for nucleic acid extraction on the same day.

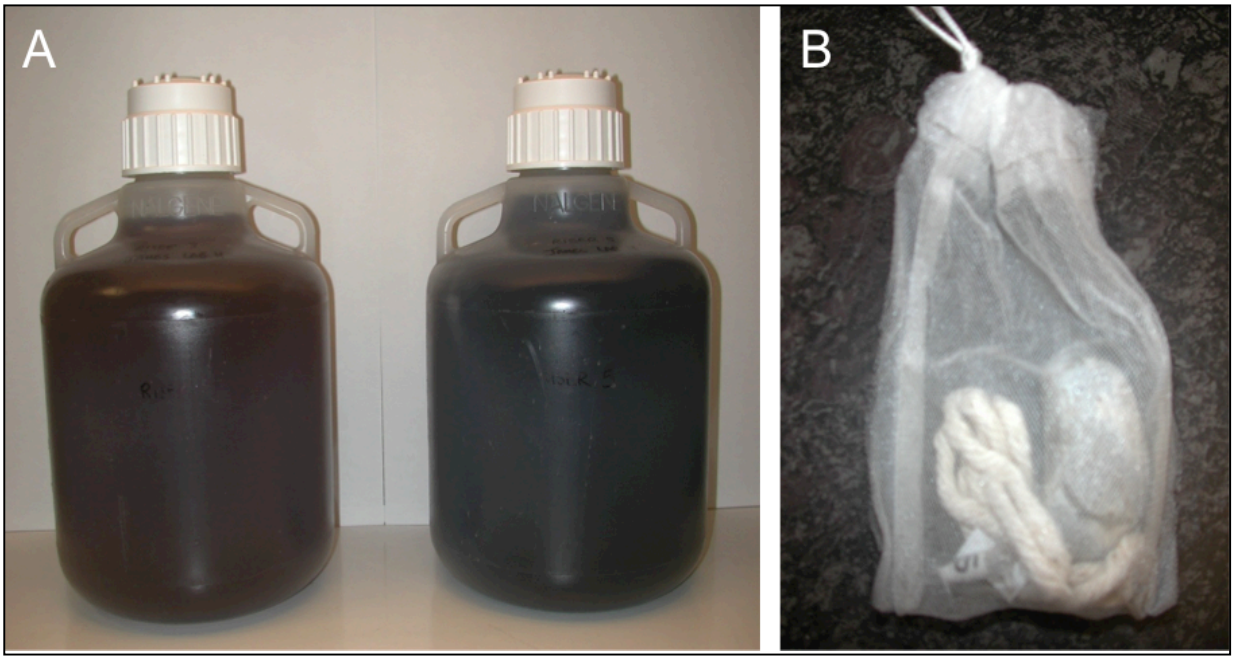


Figure 2.1 Sample collection and enrichment. Landfill leachate was stored in the laboratory at room temperature in carboys (A). Dewaxed cotton string was placed in nylon bags (B) and incubated in the leachate to allow for colonisation by cellulolytic microbes, and served as the source of nucleic acids.

Figure 2.2 A ‘riser’ sampling point at Bromborough Dock (A) and the hyperalkaline pond in Harpur Hill (B). An abundance of degrading plant matter can be seen (arrowed) near the water-sediment interface.

2.2 Nucleic acid extraction and purification

Nucleic acids were extracted from landfill leachate and hyperalkaline sediment, or from colonised cotton incubated in landfill leachate and hyperalkaline sediment using a variety of methods that are outlined below.

2.2.1 Preparation of RNase-free reagents and equipment

Reagents used for extracting both DNA and RNA were treated with diethyl polycarbonate (DEPC) in order to inactivate RNases and DNases to prevent enzymatic degradation of the samples. All reagents and glassware were rendered RNase-free by the addition of 0.1% v/v DEPC followed by overnight incubation at 37^{o}C and subsequent sterilisation by autoclaving. All plasticware was autoclaved prior to use in containers that were treated with RNaseZap solution (Ambion) and rinsed with DEPC-treated water.

2.2.2 Co-extraction of DNA and RNA from environmental samples (Griffiths *et al*., 2000)

DNA and RNA were co-extracted from 0.5 ml concentrated landfill leachate or from 0.2 g (wet weight) colonised string by adding the sample to a 2 ml screw-cap tube (Sarstedt) containing 0.5 g sterile, acid-washed glass beads (0.1 mm diameter, Sigma Aldrich). The tube was shaken at 3,400 rpm for 30 seconds in a PowerLyzer bench top homogenizer (MO BIO) following the addition of 0.5 ml 5% (w/v) CTAB buffer (equal volume 10% (w/v) cetrimonium bromide and 240 mM potassium phosphate buffer) and 0.5 ml phenol:chloroform:isoamylalcohol (25:24:1) (Sigma Aldrich). The mixture was centrifuged at 17,200 *g* for 5 minutes to separate the nucleic acids, which were dissolved in the top aqueous layer. An equal volume of chloroform:isoamylalcohol

(24:1) (Sigma Aldrich) was added, mixed and the top aqueous layer was once again extracted following centrifugation at 17,200 *g* for 5 minutes. 2 volumes PEG solution (30% (w/v) polyethylene glycol 6000 in 1.6 M NaCl) was added and the sample was subsequently incubated overnight at 4ºC to precipitate the nucleic acids. The sample was centrifuged at 17,200 *g* for 30 minutes at 4ºC and the nucleic acid pellet washed with 70% (v/v) ethanol, followed by resuspension in 50 µl RNase free water. The DNA was quantified using a Qubit (Invitrogen), and the purity was assessed using a NanoDrop 2000 (Thermo Scientific).

2.2.3 Co-extraction of DNA and RNA from pH 11.0-13.0 sediment and culture mesocosms

DNA was co-extracted with RNA from 2.0 g (wet weight) colonised cotton or from 5 ml sediment. The sample was combined with 10 ml extraction buffer (1 x phosphate buffer saline (PBS) (pH 7.0) and 0.1% (v/v) Tween 20) in a 50 ml centrifuge tube (Greiner) before vortex mixing at 2,500 rpm for 90 seconds. The sample was centrifuged at 900 *g* for 3 minutes and the supernatant containing the cellular biomass was collected. The cell suspension was then centrifuged at 14,000 *g* for 2 minutes to pellet the cells. Following the removal of the supernatant by careful aspiration and resuspension of the cell pellet in an equal volume of 1 x PBS (pH 4.0), the cells were pelleted again by centrifugation at 14,000 *g* for 2 minutes. The low pH 1 x PBS wash was performed twice when extracting RNA from colonised cotton sediment and thrice when extracting RNA from the concentrated culture mesocosm run-off. The supernatant was removed by careful aspiration and the cells were subsequently resuspended in 200 µl of 1 x PBS (pH 7.0), and the nucleic acids were then extracted using the method previously described by Griffiths *et al*. (2000).

2.2.4 High molecular weight DNA extraction method (Neufeld *et al*., 2007)

Gentle extraction of high molecular weight (HMW) DNA was carried out by the addition of 1.6 ml SET buffer (50 mM Tris-HCl (pH 9.0), 40 mM EDTA (pH 8.0) and 0.75 M sucrose), 180 µl freshly prepared lysozyme solution (9 mg ml^{-1} in 10 mM Tris-HCl (pH 8.0)) and 0.5 ml 5% CTAB buffer (equal volume 10% cetrimonium bromide and 240 mM potassium phosphate buffer) to 1 g (wet weight) colonised string in a 15 ml centrifuge tube (Greiner). The sample was incubated at 37ºC for 30 minutes, followed by the addition of 55 µl proteinase K solution (20 mg ml^{-1} in 50 mM Tris-HCl (pH 8.0)) and 200 µl 10% (w/v) sodium dodecyl sulphate, and a further incubation at 55ºC for 120 minutes was carried out. All incubations were carried out in a low speed-shaking incubator at 30 rpm. 15 ml phase lock gel (PLG) tubes (5 PRIME) were prepared by centrifugation at 1,500 *g* for 2 minutes. The string was removed and washed with 0.5 ml SET buffer, and the rinse buffer was combined with the previously collected lysate in a previously centrifuged 15 ml PLG tube. Equal volume phenol:chloroform:isoamylalcohol (25:24:1) was mixed with the combined lysate in the PLG tube by gentle inversion, followed by centrifugation at 1,500 *g* for 5 minutes. This extraction step was carried out twice. The resultant aqueous phase was combined with 2-3 volumes absolute ethanol along with 0.1 volume 3M sodium acetate (pH 5.2) and 0.02 volume glycogen (5 mg ml^{-1}), and the nucleic acids were precipitated by overnight incubation at -20^{o}C. The sample was centrifuged at 18,000 *g* for 30 minutes at 4ºC and the nucleic acid pellet was washed with 70% (v/v) ethanol, followed by resuspension in 100 µl ddH_2O.

2.2.5 High molecular weight DNA extraction method modified from Neufeld *et al*. (2007)

Gentle extraction of HMW DNA was carried out by the addition of 5 ml lysis buffer (50 mM Tris-HCl (pH 8.0), 50 mM EDTA (pH 8.0) and 1% (w/v) sodium dodecyl sulphate) and 200 µl freshly prepared lysozyme solution (9 mg ml^{-1} in 10 mM Tris-HCl (pH 8.0)) to 2 g (wet weight) colonised string in a 15 ml centrifuge tube (Greiner). The sample was incubated at 37ºC for 30 minutes, followed by the addition of 100 µl proteinase K solution (20 mg ml^{-1} in 50 mM Tris-HCl (pH 8.0)) and a further incubation at 55ºC for 120 minutes. 0.2 volume 5% (w/v) CTAB buffer (equal volume 10% (w/v) cetrimonium bromide and 240 mM potassium phosphate buffer) was added to the sample and mixed by gentle inversion, followed by incubation at 68ºC for 15 minutes. All incubations were carried out in a low speed-shaking incubator at 30 rpm. The string was removed and the remaining lysate mixed with equal volume phenol:chloroform:isoamylalcohol (25:24:1) by gentle inversion. The mixture was centrifuged at 17,200 *g* for 10 minutes to separate the nucleic acids, which were dissolved in the top aqueous layer. Equal volume chloroform:isoamylalcohol (24:1) was mixed and the top aqueous layer once again extracted following centrifugation at 17,200 *g* for 10 minutes. This extraction step was carried out twice. 2 volumes PEG solution (30% (w/v) polyethylene glycol 6000 in 1.6 M NaCl) was added and the sample was subsequently incubated overnight at 4ºC to precipitate the nucleic acids. The sample was centrifuged at 17,200 *g* for 30 minutes at 4ºC and the nucleic acid pellet was washed with 70% (v/v) ethanol, followed by resuspension in 200 µl ddH_2O.

2.2.6 High molecular weight DNA isolation using the Meta-G-Nome DNA Isolation Kit (Epicentre)

Briefly, fresh extraction buffer was made by supplementing the extraction buffer provided with the kit by addition of 0.1% (v/v) Tween 20 (Sigma-Aldrich). The colonised string was combined with 10 ml extraction buffer in a 50 ml centrifuge tube (Greiner) before vortex mixing at 2,500 rpm for 60 seconds. The sample was centrifuged at 1,600 *g* for 4 minutes and the supernatant containing the cellular biomass was collected. A further 20 ml of extraction buffer was added to the pelleted matter and vortex mixed at 2,500 rpm for 60 seconds, before being centrifuged at 900 *g* for 3 minutes. Following collection of the supernatant and pooling with previously collected supernatant; the aforementioned step was repeated to yield 50 ml of cell suspension. This was passed through a 0.45 μm filter membrane to trap the cellular biomass, which was dislodged into 1.5 ml filter wash buffer by vortex mixing at 2,500 rpm for 2 minutes in a 50 ml centrifuge tube. The cell suspension was centrifuged at 14,000 *g* for 2 minutes to pellet the cells, and the cell pellet was resuspended in a combination of 100 μl 5% CTAB buffer (equal volume 10% cetrimonium bromide and 240 mM potassium phosphate buffer) and 200 μl TE buffer (10 mM Tris-HCl (pH 7.5), 1 mM EDTA).

The cell suspension was gently mixed with 2 μl Ready-Lyse Lysozyme and 1 μl RNase A solutions provided with the kit, followed by incubation at 37ºC for 30 minutes. A further incubation was carried out for 15 minutes at 65ºC after gently mixing the sample with 300 μl Meta-Lysis Solution (2x) and 1 μl Proteinase K provided with the kit. The sample was incubated on ice for 5 minutes before 350 μl MPC Protein Precipitation Reagent was added and the cell debris was pelleted by centrifugation at 20,000 *g* for 10 minutes at 4ºC. The supernatant was incubated on ice for 15 minutes after the addition of 570 μl absolute propan-2-ol and mixing by inversion, followed by centrifugation at

20,000 *g* for 15 minutes at 4ºC to pellet the DNA. Propan-2-ol was removed by careful aspiration and the DNA pellet was washed with 70% (v/v) ethanol, followed by resuspension in 50 µl TE buffer.

2.2.7 RNA removal from DNA and RNA co-extracts

DNA that was co-extracted with RNA from colonised cotton, landfill leachate, hyperalkaline sediment or concentrated mesocosm run-off was treated with RNase A (Invitrogen) for RNA removal. The samples were made up to 150 µl with nuclease-free water and RNase A was added to a final concentration of 100 µg ml^{-1} before being incubated at 37^{o}C for 15 minutes. Purification of samples following nuclease treatment was performed by brief vortex mixing following the addition of 200 µl phenol:chloroform:isoamylalcohol (25:24:1) (Sigma Aldrich). The mixture was centrifuged at 17,200 *g* for 5 minutes to separate the nucleic acids, which were dissolved in the top aqueous layer. Equal volume chloroform:isoamylalcohol (24:1) (Sigma Aldrich) was mixed and the top aqueous layer was once again extracted following centrifugation at 17,200 *g* for 5 minutes. The nucleic acids were precipitated with two volumes 100% ethanol and 0.25 volumes 10M ammonium acetate, followed by incubation for 30 minutes at -80^{o}C. The sample was centrifuged at 18,000 *g* for 30 minutes at 4ºC and the nucleic acid pellet was washed with 70% (v/v) ethanol, followed by resuspension in 30 µl ddH_2O.

2.2.8 mRNA extraction, purification and amplification

Total RNA was co-extracted with DNA from 0.2 g (wet weight) colonised cotton incubated in landfill (section 2.2.2), or from 2.0 g colonised cotton or 5 ml hyperalkaline sediment (section 2.2.3) using the methods previously described, with

the following modifications: 100 µl of 14.3 M 2-mercaptoethanol (Sigma-Aldrich) was added to each extraction tube before bead-beating and the nucleic acids were precipitated by overnight incubation at -20^{o}C in 2-3 volumes absolute ethanol along with 0.1 volume 3M sodium acetate (pH 5.2) and 0.02 volume glycogen (5 mg ml^{-1}) after the addition of 5 µl RNasin Plus RNase inhibitor (Promega). All surfaces, gloves and pipette exteriors were treated with RNase Zap solution before commencing the extraction and handling of RNA. RNase-free water and filter pipette tips were used throughout.

2.2.8.1 DNA removal from RNA and DNA co-extracts

DNA was removed from the total nucleic acids (TNA) using 2 units of Turbo DNase (Invitrogen) per µg of DNA in a sample followed by incubation of the samples at 37^{o}C for 30 minutes. This was followed by the addition of 250 µl of phenol:chloroform:isoamylalcohol (25:24:1) (Sigma Aldrich) and brief vortex mixing of the sample. The mixture was centrifuged at 17,200 *g* for 5 minutes to separate the nucleic acids, which were dissolved in the top aqueous layer. Equal volume chloroform:isoamylalcohol (24:1) (Sigma Aldrich) was mixed and the top aqueous layer was once again extracted following centrifugation at 17,200 *g* for 5 minutes. The nucleic acids were precipitated with two volumes 100% ethanol and 0.25 volumes 10M ammonium acetate, followed by incubation for 30 minutes at -80^{o}C. The sample was centrifuged at 18,000 *g* for 30 minutes at 4ºC and the nucleic acid pellet was washed with 70% (v/v) ethanol, followed by resuspension in 30 µl RNase-free H_2O.

2.2.8.2 Removal of small RNAs

Small RNAs, including transfer RNA (tRNA) and 5S ribosomal RNA (rRNA), were removed from total RNA using the silica column-based RNeasy MinElute Clean-up kit (Qiagen) following the manufacturer's protocol. This procedure also served to remove residual impurities like DNA degradation products and excess salt.

2.2.8.3 rRNA depletion using MICROB*Express* Bacterial mRNA Enrichment kit (Ambion)

Bacterial SSU rRNA removal was performed using the MICROB*Express* Bacterial mRNA Enrichment kit (Ambion) following the manufacturer's protocol. MICROB*Express* kit employs a capture hybridisation approach using pre-designed probes attached to Oligo MagBeads. The probes hybridised to known 16S and 23S rRNA in the sample, which were subsequently separated from the rest of the RNA by virtue of the Oligo MagBeads being removed using a Magna-Sep magnetic eppendorf stand (Invitrogen).

2.2.8.4 rRNA removal using Terminator 5'-Phosphate-Dependent Exonuclease (Epicentre)

As instructed by the manufacturer, each sample was treated with 2 units of Terminator 5'-Phosphate-Dependent Exonuclease used with 0.1 volume 10X reaction buffer B for further bacterial rRNA depletion by incubation at 42^{o}C for 30 minutes. The reaction was terminated by the addition of 1 µl of 100 mM EDTA. The RNeasy MinElute Clean-up kit was used to remove excess salts and degradation products following the manufacturer's protocol.

2.2.8.5 Poly-A tailing of mRNA

Polyadenylation of the enriched mRNA was performed using *E. coli* Poly(A) Polymerase enzyme (New England Biolabs) following the manufacturer's protocol. Briefly, 10 units of enzyme were used in each 20 µl reaction, followed by incubation for 30 minutes at 37ºC. The reaction was terminated by addition of EDTA at a final concentration of 10 mM into the reaction buffer.

2.2.8.6 mRNA amplification using MessageAmp II aRNA Amplification kit (Invitrogen)

The polyadenylated mRNA was amplified using MessageAmp II aRNA Amplification kit (Invitrogen) according to the manufacturer's protocol, with the following modifications: the T7 oligo(dT) primer provided with the kit was replaced with a primer that contained the BpmI restriction site (T7 BpmI oligo(dT) primer), and the *in vitro* transcription was performed in a hybridisation oven at 37^{o}C with an increased incubation time of 18 hours. The protocol relies on reverse transcription of the mRNA into double strand complementary DNA (ds cDNA) followed by *in vitro* transcription to synthesise antisense RNA (aRNA), therefore producing an amplification effect.

2.3 Qualitative and quantitative assessment of nucleic acids

Reliable quantification of both DNA and RNA was obtained using the Qubit fluorometer platform (Invitrogen). NanoDrop 2000 spectrophotometer (Thermo Scientific) readings were taken to assess purity of samples, based on the ratios of the absorbance values at 280 and 260 nm (A_{260}/A_{280}) to provide an indication for chemical and/or protein contamination. Agarose gel electrophoresis was performed for visual

inspection of shearing and enzymatic degradation, as evidenced by poor band integrity. Integrity of the RNA was more reliably determined using the prokaryote total RNA 6000 (nano) assay on a 2100 Bioanalyzer (Agilent).

2.4 Anaerobic fungal culturing method

A modified version of the Hungate technique (Hungate, 1950) for anaerobic culturing was developed to isolate anaerobic cellulolytic fungi. Briefly, 50 mg yeast extract, inorganic salts solution at 1 x concentration (Table 2.1) and 0.0001% resazurin dye (w/v) final concentration were mixed and boiled for 15 minutes. Resazurin was used as a redox indicator, as it turns from dark blue to colourless when the conditions change from aerobic to anaerobic. 10 ml of the boiled solution was dispensed into 22 ml glass culture tubes (Supelco) containing 0.4 g of cellulosic substrate: powdered cellulose (Avicel), cellobiose, carboxymethyl cellulose (CMC), strips of Whatman no. 1 filter paper, dewaxed cotton string or hay. Following sterilisation, a reducing agent combination of filter-sterilised cysteine hydrochloride (0.005% w/v), sodium pyruvate (0.01% w/v) and sodium thioglycolate (0.01% w/v) was added to each tube, along with filter-sterilised chloramphenicol, streptomycin and ampicillin (0.007% w/v each). Anaerobic gas mix (BOC), consisting of 5% CO_2, 10% H_2 and 85% N_2, was bubbled through the media to drive out any residual oxygen and was also used to fill the head-space of the culture tubes, which had rubber septa (Supelco) between the tube and a screw-on lid to ensure that the cultures remained airtight during incubation. All reducing agent solutions were also saturated with the anaerobic gas mix and were stored in airtight 22 ml glass culture tubes until use to prevent their oxidation.

The tubes were set up in triplicate for each treatment and were incubated overnight at 37^{o}C in an anaerobic cabinet (Don Whitley) to allow the culture media to

be completely reduced. Landfill leachate was added as inoculum to each tube to a final volume of 20 ml and incubated for 8-12 weeks in the anaerobic chamber at 37^{o}C. Sub-culturing was performed by the addition of 2 ml of previously incubated culture as inoculum into 18 ml of inorganic media as described above. PCR analysis was performed for detection of anaerobic fungi following incubation.

2.5 Preparation, replication and storage of an environmental fosmid library using high molecular weight DNA

High molecular weight DNA extracted from colonised string incubated in landfill leachate using either the method modified from Neufeld *et al*. (2007) or the Meta-G-Nome DNA Isolation Kit (Epicentre) was selected to prepare an environmental fosmid library using the CopyControl Fosmid Library Production Kit (Epicentre) following manufacturer's protocol. Briefly, the metagenomic DNA was end-repaired to generate blunt-ended, 5'-phosphorylated DNA. The end-repaired DNA was then size selected by pulsed field gel electrophoresis (PFGE) for 18 hours using a 1% (w/v) low melting point (LMP) agarose gel. The 30kb-50kb DNA fraction was extracted from the gel and ligated into CopyControl pCC1FOS fosmid vector. The ligation product was then packaged into a MaxPlax lambda packaging extract, which was subsequently used to infect an *E. coli* EPI300-T1^{R} plating strain. The titre of the packaged fosmid clones was determined and the fosmid library was plated out on LB agar plates containing 12.5 μg/ml chloramphenicol as a selection marker. Following growth for 18 hours, the fosmid clones were picked and propagated in freezing medium (Table 2.2) in 2 ml MasterBlock 96-well plates (Greiner), with each well consisting of a pool of ~ 500 fosmid clones. The fosmid library was replicated into further 2 ml 96-well plates and stored at -80ºC until further use.

2.6 Assays for screening of environmental fosmid libraries

The assays used to screen the fosmid libraries for the presence of cellulases and xylanases are outlined below.

2.6.1 Congo red assay

The surface of LB agar plates, supplemented with 0.2% (w/v) carboxymethylcellulose (CMC) and 12.5 μg ml^{-1} chloramphenicol, was covered with 82 mm circular 0.45 μm filter membranes (Pall) before a 10^{-4} or 10^{-5} dilution of an induced fosmid clone culture was plated onto them. Following overnight incubation at 37ºC, the orientation of the membrane that contained the bacterial colonies was marked onto the plate before removal of the membrane. The agar plates were subsequently flooded with 0.1% Congo red solution and shaken at 25 rpm for 30 minutes to allow for even staining of the CMC. The Congo red solution was washed off with 1M NaCl solution by shaking at 25 rpm for 15 minutes twice. The plates were than visualised using a light box to detect any zones of clearance around the colonies against the red background of the medium. Marked orientation of the plates was used to determine the colonies on the membrane corresponding to the zones of clearance on the plate, and positive clones were propagated overnight in fresh freezing medium (Table 2.2).

Table 2.1 Composition of 1 x inorganic salts solution used for anaerobic culturing

Inorganic salt	**Final concentration (%) (w/v)**
$(NH_4)_2SO_4$	0.05
K_2HPO_4	0.5
KH_2PO_4	0.2
$CaCl_2$	0.005
$MgSO_4$	0.005
NaCl	0.1

Table 2.2 Composition of the freezing medium used for propagating fosmid clones

Component	**Concentration (%) (w/v)**
K_2HPO_4	0.63
KH_2PO_4	0.18
$Na_3C_6H_50_7$	0.05
$MgSO_4.7H_2O$	0.01
$(NH_4)_2SO_4$	0.09
LB medium	2.0
Glycerol	4.4*

* measured in (v/v)

2.6.2 *p*-nitrophenyl β-D-cellobioside (pNPC) assay

The broth based pNPC assay was performed in 2 ml MasterBlock 96-well plates (Greiner) to detect cellobiohydrolase activity. Briefly, 100 µl of induced fosmid clones grown in freezing medium (Table 2.2) was added to 200 µl of 50 mM ammonium acetate (pH 5.0) supplemented with 5 mM pNPC substrate. The assay mix was incubated overnight at 37ºC before 700 µl of 1M Na_2CO_3 was added to further develop the yellow colour from the release of *p*-nitrophenol (pNP). The assay mix was centrifuged at 14,000 *g* for 5 minutes to pellet the cellular biomass and 200 µl of the supernatant from each well was transferred to a new 0.2 ml 96-well plate (Greiner). Spectrophotometry was subsequently performed to determine the absorbance from each well at 410 nm.

2.6.3 Broth based AZCL-HE-Cellulose and AZCL-Xylan assays

The broth based AZCL-HE-Cellulose and AZCL-Xylan assays were performed in 1 ml MasterBlock 96-well plates (Greiner). Briefly, 200 µl of induced fosmid clones grown in freezing medium (Table 2.2) was added to 200 µl of 200 mM sodium acetate (pH 5.0) supplemented with 2% (w/v) of the appropriate dyed substrate. The assay mix was incubated overnight at 37ºC and centrifuged at 14,000 *g* for 5 minutes thereafter to pellet the cellular biomass along with the particulate substrate. 200 µl of the supernatant from each well was transferred to a new 0.2 ml 96-well plate (Greiner). Spectrophotometry was subsequently performed to determine the absorbance from each well at 595 nm to monitor the release of azure dye.

2.6.4 Agar based AZCL-HE-Cellulose and AZCL-Xylan assays

LB agar plates, supplemented with 12.5 μg ml^{-1} chloramphenicol, were prepared before a 10^{-4} or 10^{-5} dilution of an induced fosmid clone culture was plated onto them. Following overnight incubation at 37ºC, the surface of the plates was gently covered with 0.7% (w/v) agar supplemented with 0.05% (w/v) of the appropriate dyed substrate before being incubated overnight at 37ºC. The particulate substrate was cleaved to release blue dye in the top agar surrounding any positive clones.

2.7 End point PCR

PCR was carried out using either Biomix Red (Bioline) or Phusion high-fidelity DNA polymerase (Finnzymes).

2.7.1 Biomix Red

PCR reactions were performed in a total volume of 25 μl. The reaction mix consisted of the following components: 12.5 μl of 2X Biomix Red, 50-80 ng of DNA template, 0.4 mM each of the forward and reverse primer, made up to 25 μl with sddH_2O. The PCR cycling conditions were as follows: initial denaturation at 94^oC for 5 minutes; 35 cycles of denaturation at 94^oC for 1 minute, annealing at the specific temperature for the primer set for 1 minute and extension at 72^oC for 1.5 minutes; and final extension at 72^oC for 10 minutes. Conditions were altered as necessary to improve amplification and specificity.

2.7.2 Phusion

Reactions requiring high sensitivity and fidelity were performed using Phusion high-fidelity DNA polymerase (Finnzymes), which displays a low error rate due to its

proof reading activity. Reactions were carried out in 25 µl volumes. The reaction mix was made up as follows: 0.2 mM each of the forward and reverse primer, 0.2 mM each dNTP, 12.5 µl of 2 x Phusion HF buffer, 1 µl of Phusion polymerase, 50-80 ng of DNA template, made up to 25 µl with sddH_2O. Typical PCR cycle conditions were as follows: initial denaturation at 98^oC for 45 seconds; 35 cycles of 98^oC for 10 seconds, 30 seconds at the appropriate annealing temperature for the primer set, 72^oC for 30 seconds; and a final extension at 72^oC for 5 minutes. In some cases, Phusion reactions were performed in a 50µl volume if yield was poor, or to boost yields obtained from target DNA of low quantity.

2.7.3 Nested PCR

Nested PCR was used to improve the sensitivity of detection when DNA from difficult environmental samples was to be amplified, with issues arising due to inadequate template DNA to provide first round amplification products. Where nested PCR was performed, the amplification products from the first round of PCR were diluted 100-fold and re-amplified using a specific set of primers which target a section of the DNA fragment that is internal to the amplification products from the first round of PCR.

2.7.4 PCR primers

Table 2.3 lists the PCR primer sets used in this study, their purpose and general characteristics. These primers have either been obtained from information in previous publications or have been provided with commercial kits. The specific conditions of use of these primers are discussed in the appropriate results chapters.

Table 2.3 Details of the primers used in this study

Primer set	Purpose	Sequence 5'-3'	Target	Annealing temperature	Amplicon size (bp)	Reference
pA pH	End point PCR	AGAGTTTGATCCTGGCTCAG AAGGAGGTGATCCAGCCGCA	General bacteria	55ºC	~1530	Edwards *et al*., 1989
Cther 650 Cther 1352	End point PCR	TCTTGAGTGYYGGAGAGGAAAGC GRCAGTATDCTGACCTRCC	*Clostridium* Group III	60ºC	~720	Van Dyke & McCarthy, 2002
FIB 1F FIB 2AR	End point PCR	CCGKSCCAACGSSCGG ATCTCTCGCYGCGGCGWTYCC	*Fibrobacter*	60ºC	~855	McDonald *et al*., 2008
NS1-Euk Univ 1390	End point PCR	CCAGTAGTCATATGCTTGTC GACGGGCGGTGTGTACAA	General Eukarya	50ºC	~1600	Suzuki *et al*., 2000
Chyt 719F Chyt 1553R	End point PCR	GCACTTCATTGTGTGTACTG GGATGAAACTCGTTGACTTC	*Neocallimastigales*	60ºC	~787	Lockhart *et al*., 2006
M 13F M 13R	Molecular cloning	GTAAAACGACGGCCAGT GCGGATAACAATTTCACACAGG	pGEM-T Easy Vector	60ºC	N/A	N/A
T7-BpmI-(dT)16VN	Production of first-strand cDNA from polyA+ RNA	TAATACGACTCACTATAGGGAGA *GACCTC*$(TTTT)_4$VN	Non-specific for polyA+ RNA	N/A	N/A	Included with the MessageAmpII kit (Ambion)

2.8 DNA clean up

To remove residual contamination from DNA preparations, or to clean up DNA after PCR and other enzymatic treatments, DNA was cleaned up using the spin column based method of ISOLATE II PCR and Gel Kit (Bioline). Clean-up was performed according to the manufacturer's instructions, with the following modification: the DNA was eluted from the spin columns using ddH_2O, following two elutions of 10 µl each. The use of the elution buffer provided with the kit was avoided to prevent interference with downstream applications.

2.9 Agarose gel electrophoresis

Agarose gel electrophoresis was performed to visualise nucleic acid extracts and PCR amplification products. Gels comprised 1% (w/v) agarose in 1 x TAE (Tris-acetate EDTA) buffer diluted with ddH_2O from a 50 x stock of TAE (2M Tris; 57.1 ml L^{-1} glacial acetic acid; 0.05M EDTA (pH 8.0). Agarose gels were stained with ethidium bromide added to a concentration of 0.5 µg ml^{-1}. The electrophoresis was performed in 1 x TAE buffer at 110 V for 60 minutes. The gel tank was cleaned using RNase*Zap* (Invitrogen) and the TAE running buffer was replaced before performing gel electrophoresis on RNA samples. Nucleic acid fragment size was determined by comparison to appropriate molecular size markers, including Hyperladder 1 (Bioline) or GeneRuler™ 100 bp DNA Ladder Plus (Thermo Scientific). All nucleic acids were visualised using ultra-violet (UV) trans-illumination on a Gene Genius bio-imaging system (Syngene).

2.9.1 Pulsed field gel electrophoresis

PFGE was performed to determine the size of the high molecular weight DNA by running a 1% (w/v) agarose gel in 1 x TAE buffer in a CHEF-DR III Pulsed Field Electrophoresis Systems machine (Bio-Rad). GeneRuler^{TM} high range DNA ladder (Thermo Scientific) was used the molecular size marker. The gel electrophoresis conditions were as follows: 14^{o}C, 1-6 seconds switch time, 6 V cm^{-1} voltage, 120^{o} angle and 16-18 hours run time. Agarose gels were stained with ethidium bromide added to a concentration of 0.5 μg ml^{-1}.

2.9.2 Extraction of DNA bands from agarose gels

DNA bands of interest were purified from LMP agarose gels using the ISOLATE II PCR and Gel Kit (Bioline) following the manufacturer's instructions, with the following modification: the DNA was eluted from the spin columns using ddH_2O, following two elutions of 10 μl each.

2.10 Molecular cloning of rRNA gene PCR amplification products

PCR products of interest were size-selected by gel electrophoresis at 110 V for 60 min using a 1% (w/v) LMP agarose (Melford) gel. Following visualisation under long wave ultraviolet light (365 nm), the DNA bands of interest were excised using a sterile scalpel blade and purified using the ISOLATE II PCR and Gel kit (Bioline). The amplified rDNA was cloned into a pGEM-T Easy vector (Promega), followed by transformation into *E. coli* JM109 high efficiency chemically competent cells according to the manufacturer's protocol. Subsequently, blue-white screening was performed to isolate successful transformants. Individual white colonies were used to inoculate Luria broth (LB), supplemented with ampicillin (100 μg ml^{-1} w/v), and propagated overnight

at 37^{o}C. Plasmid DNA was extracted from overnight cultures of the transformants using the ISOLATE II Plasmid Mini kit (Bioline) following the manufacturer's protocol and EcoRI (New England Biolabs) restriction digestion was performed for 60 minutes at 37ºC followed by gel electrophoresis to ensure that the insert in each plasmid was of the desired size. Isolated plasmid DNA was subsequently sent for sequencing in both directions to GATC Biotech, Germany.

2.11 Sequence data analysis

Plasmids incorporating rDNA PCR amplification products were Sanger sequenced following molecular cloning, and taxonomic classification of the sequences was performed by alignment against the NCBI nr nucleotide database using BLAST (Altschul *et al*., 1990) with an E-value cut-off of 0.001 (chapter 3, section 3.4.1). 454 pyrosequenced rDNA amplicons were subjected to quality control, sequence denoising and Operational Taxonomic Unit (OTU) picking against the SILVA database (Pruesse *et al*., 2007) using Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso *et al*., 2010) with an E-value cut-off of 0.001, as detailed in chapter 3 (section 3.4.2). Data pre-processing and quality control of metagenomic and metatranscriptomic reads generated by Illumina MiSeq high throughput sequencing was performed using Galaxy (Afgan *et al*., 2016, accessible at www.usegalaxy.org). Taxonomic and functional analyses of the high throughput sequence data was performed using the MG-RAST webserver (Meyer *et al*., 2008) (chapter 4, section 4.3.2). Open reading frame (ORF) prediction was carried out using the MetaGeneMark webserver (Zhu *et al*., 2010), and those were assigned to glycoside hydrolase families by querying against the protein family database Pfam A (Finn *et al*., 2010) with an E-value cut-off of 0.001, as detailed in chapter 4 (section 4.3.4). Illumina MiSeq-sequenced fosmid DNA reads were

processed using Galaxy webserver and assembled using Geneious version 7.1 (www.geneious.com, Kearse *et al*., 2012). The taxonomic identity of ORFs containing glycoside hydrolase domains was elucidated by a BlastX search against the NCBI nr database, following ORF prediction using MetaGeneMark and assignment of those to glycoside hydrolase families by querying against the Pfam database (chapter 5, section 5.4.5).

Chapter 3

Detection and enrichment of obligately anaerobic cellulolytic fungi of the order *Neocallimastigales* from landfill leachate

3.1 Background

While the majority of microbially mediated cellulose degradation occurs under aerobic conditions, anaerobic cellulose degraders are also responsible for maintaining the flow of environmentally significant amounts of carbon in the biosphere. Much of our current understanding of anaerobic cellulose degradation arises from studies conducted on the herbivore gut (Rubin, 2008). Microbes such as *Neocallimastix frontalis*, *Clostridium thermocellum* and *Fibrobacter succinogenes*, amongst others, perform efficient hydrolysis of cellulose and due to the various energy constraints imposed under anoxic conditions, their enzymes demonstrate much greater specific activity compared to their aerobic counterparts (Lynd *et al*., 2002). Early research by Wilson & Wood (1992a, b) suggested that anaerobic chytrid fungi are amongst the most potent cellulose degraders in the biological world, a feat possible not only because of their filamentous morphology that readily allows access to cellulosic biomass, but also due to the synergistic action of their cellulosomes coupled with a rich repertoire of lignocellulosic enzymes (Leschine, 1995).

Orpin (1975) first brought to light the existence of anaerobic fungi in the sheep rumen, as the motile stage of *Neocallimastix frontalis* was previously assumed to be the protozoan flagellate *Callamastix frontalis*. Numerous reports relating to techniques involved in their isolation and characterisation have since been published (Bauchop, 1979; Joblin, 1981; Lowe *et al*., 1985; Lowe *et al*., 1987). Members of

Neocallimastigales, the order of fungi belonging to the anaerobic cellulolytic class *Chytridiomycetes,* have subsequently been isolated from the digestive tract of a variety of herbivores from numerous geographical locations (Ljungdahl, 2008). To date, representatives of the order *Neocallimastigales* have not been isolated from any environment other than the herbivore gut, and it was assumed that the presence of these fungi was restricted to such an environment (Haitjema *et al*., 2014). There is however molecular evidence for the presence of these fungi in landfill sites through the targeting of the 18S rRNA gene (Lockhart *et al*., 2006; McDonald *et al*., 2012), leading to the hypothesis that environmental relatives of the ruminant microbe *Neocallimastigales* could be cultured from other anoxic environments, including landfill leachate. This suggests a wider ecological distribution of these fungi in anoxic environments and a potentially crucial role in cellulose hydrolysis in general.

It is important that attempts to isolate these anaerobic fungi are made to better understand their physiology, and to be able to perform whole genome sequencing and transcriptomic analysis. This would facilitate research into the conceivably remarkable portfolio of enzymes they possess for lignocellulose hydrolysis and their potential industrial application in consolidated biomass processing, as only one case of genome analysis has been reported thus far (Youssef *et al*., 2013).

3.2 Detection of cellulolytic microbes using PCR

Municipal waste landfill sites at Bromborough Dock and Bidston Moss (Wirral, Northwest England) were sampled, and landfill leachate was collected in 10 L carboys (Nalgene) as described in section 2.1.2. Dewaxed cotton string (generated as described in section 2.1.1) was subsequently incubated in leachate samples BD1, BD2 and BD3 from Bromborough Dock, and samples BM3E, BM3F, BM3G, BM3H and BM1J from

Bidston Moss, followed by incubation at room temperature. After an incubation period of 3 months, the incubated string was assessed for signs of degradation and microbial biofilm formation visually. A PCR-based investigation was carried out into assessing the biodiversity of the leachate samples and the biofilm formed on the string. While the string incubated in carboys BD1, BD2, BD3, BM3E and BM1J demonstrated clear signs of thick biofilm formation, this result was not observed for string incubated in carboys BM3F, BM3G and BM3H. Consequently, string from those 3 samples was omitted from the PCR analysis. Apart from direct and nested PCR detection of *Neocallimastigales* fungi, presence of certain cellulolytic bacteria was also examined using direct and nested PCR. *Fibrobacter* and *Clostridium* cluster III (Collins *et al*., 1994) bacteria have been reported to play a significant role in anaerobic cellulose decomposition, while their presence in landfill sites has been confirmed through molecular biological approaches (Van Dyke & McCarthy, 2002; McDonald *et al*., 2008).

Direct PCR amplification of the 16S rRNA gene was performed for bacteria, *Fibrobacter* and *Clostridium* cluster III representatives, and 18S rRNA gene amplification was performed for eukaryotes and *Neocallimastigales* fungi (as described in section 2.7.1) using the primers listed in Table 2.3. For direct PCR, extracted DNA was amplified using the group specific primers, *i.e.* for general bacteria, universal eukarya, *Clostridium* cluster III, *Fibrobacter* and anaerobic fungi. Where nested PCR was performed, the amplification products from the first round of PCR were diluted 100-fold before being re-amplified using the group specific primer set for *Chytridiomycete* 18S rRNA genes. Nested PCR was not needed for any bacterial amplification. Agarose gel electrophoresis was performed (as described in section 2.9) for visual detection of PCR products.

3.2.1 PCR amplification from landfill leachate

25 ml of landfill leachate sample stored in each carboy was concentrated by centrifugation at 10,000 *g* for 5 minutes, followed by resuspension of the pelleted cellular biomass and particulate matter into 5 ml leachate. DNA was then extracted from 0.5 ml each of the concentrated leachate sample using the method described by Griffiths *et al*. (2000) (section 2.2.2). The use of the CTAB buffer in the extraction method ensured that the extracted DNA was free from the large amount of humic and fulvic acids typically found in landfill leachate. The amount of DNA extracted varied between samples, with values ranging from ~ 400 ng to ~ 825 ng. The purity of DNA was assessed as the A_{260}/A_{280} ratio obtained from the NanoDrop 2000 spectrophotometer (Thermo Scientific), and readings ranged from 1.80 to 1.99. DNA samples were considered to be 'clean' and devoid of impurities such as carried over protein and phenol since the recorded A_{260}/A_{280} readings were over 1.70.

The PCR amplification results are presented in Table 3.1. Bacteria were detected in all leachate samples, whereas eukaryotic DNA was not present in any of the Bromborough Dock samples but was present in all Bidston Moss samples except BM3E. *Clostridium* cluster III representatives were found in all the leachate samples, and *Fibrobacter* spp. were detected in all leachate samples except BM3F and BM3H. This is interesting as 2 out of 3 leachate samples where incubated strings did not demonstrate signs of degradation lacked *Fibrobacter*, perhaps pointing towards the importance of these bacteria in cellulosic biomass decomposition (McDonald *et al*., 2008; McDonald *et al*., 2009; McDonald *et al*., 2012; Ransom-Jones *et al*., 2012). No anaerobic fungi were detected using direct PCR. However, a very weak positive

reaction for *Neocallimastigales* was observed from the BM3F, BM3G, BM3H and BM1J samples, but not from the BD1, BD2, BD3 or the BM3E samples. This was expected as eukaryotic DNA was not found in BD1, BD2, BD3 and BM3E leachate either.

The PCR data presented here is largely consistent with the findings of McDonald *et al*. (2012), who performed direct and nested PCR-based detection of *Clostridium* cluster III, *Fibrobacter* and anaerobic fungi in Bromborough Dock leachate and Bidston Moss leachate. Positive PCR reactions were reported for *Clostridium* cluster III in Bromborough Dock (direct and nested) and Bidston Moss (nested), for *Fibrobacter* in Bromborough Dock (direct and nested) and Bidston Moss (nested), and for anaerobic fungi in Bidston Moss (nested). The presence of anaerobic fungi was not determined using direct PCR. Nested PCR product was obtained for anaerobic fungi from only one out of five Bromborough Dock risers (riser 2), whereas no anaerobic fungi were detected from Bromborough Dock leachate in this study.

3.2.2 PCR amplification from cotton incubated in landfill leachate

DNA was also isolated from 0.5 g colonised cotton string each incubated in carboys BD1, BD2, BD3, BM3E and BM1J using the method described by Griffiths *et al*. (2000) (section 2.2.2). The total DNA yield ranged from ~ 1.3 μg to ~ 3.1 μg, with the A_{260}/A_{280} values ranging from 1.94 to 2.06. The PCR amplification was performed on suitably clean DNA samples and the results are presented in Table 3.2. Bacteria were detected from all the colonised string samples, whereas eukaryotes were only detected from the strings incubated in the Bidston Moss leachate and the combined sample comprising of BD1, BD2, BD3, BM3E and BM1J DNA. *Fibrobacter* and *Clostridium* cluster III representatives were also found in the biofilm on all five string samples.

Anaerobic fungi could not be detected using direct PCR from any sample. Using nested PCR, a very weak anaerobic fungal PCR positive reaction was observed from the two Bidston Moss samples and the combined sample, but no bands were observed from the 3 Bromborough Dock samples. Again, this was expected as no eukaryotic DNA was amplified from the extracted DNA from the Bromborough Dock samples either.

Table 3.1. Table showing the results of the PCR analysis of the DNA extracted from the landfill leachate collected from various risers at Bromborough Dock (BD) and Bidston Moss (BM) landfill sites. Direct PCR was performed unless otherwise stated.

	BD1	BD2	BD3	BM3E	BM3F	BM3G	BM3H	BM1J
General bacteria	++	++	++	++	++	++	++	++
General eukarya	-	-	-	-	++	++	++	++
Fibrobacter	++	++	++	++	-	++	-	++
***Clostridium* cluster III**	++	++	++	++	++	++	++	++
Anaerobic fungi	-	-	-	-	-	-	-	-
Anaerobic fungi (nested)	-	-	-	-	+	+	+	+

+ + Strong band on gel
+ Weak band on gel
- No PCR product detected

Table 3.2. Table showing the results of the PCR analysis of the DNA extracted from the colonised cotton samples that were incubated in the landfill leachate collected from various risers at Bromborough Dock (BD) and Bidston Moss (BM) landfill sites. Direct PCR was performed unless otherwise stated. The combined sample contained DNA from strings incubated in carboys BD1, BD2, BD3, BM3E and BM1J.

	BD1	BD2	BD3	BM3E	BM1J	Combined
General bacteria	++	++	++	++	++	++
General eukarya	-	-	-	++	++	++
Fibrobacter	++	++	++	++	++	++
***Clostridium* cluster III**	++	++	++	++	++	++
Anaerobic fungi	-	-	-	-	-	-
Anaerobic fungi (nested)	-	-	-	+	+	+

+ + Strong band on gel
+ Weak band on gel
- No PCR product detected

3.2.3 Chemical data from landfill leachate

Chemical data relating to the leachate samples collected from the risers across the two landfill sites was obtained from Merseyside Waste Disposal Authority and Biffa Waste Services Limited, the management companies in charge of Bidston Moss and Bromborough Dock landfill sites, respectively. The data are presented in Table 3.3, along with the current groundwater trigger levels for the relevant compounds as a means for comparison. High biochemical oxygen demand (BOD) and chemical oxygen demand (COD) in Bromborough Dock leachate suggests that it is very rich in organic matter, compared to the leachate in Bidston Moss boreholes 3F, 3G, 3H and 1J. BM3E was the only borehole in Bidston Moss where the sampled leachate was rich in organic content. However, this was probably because a dead animal was found in the borehole whilst sampling, and it was the only borehole that ran dry whilst collecting leachate, implying that a lot of sediment as well as decaying organic matter was present in the sample. All leachate samples were found to be rich in inorganic matter upon comparison with groundwater trigger levels that are detailed in the current legislation.

It is suggested that presence of high organic and high inorganic matter facilitates bacterial growth leading to eukaryotes being outcompeted (Jacquet *et al*., 2002), which might explain their absence in all Bromborough Dock leachate samples as well as the BM3E leachate. Adding an external carbon source like cellulose would allow specialist cellulose degraders, including anaerobic fungi, to colonise and grow on the string in BM3E. This would explain the observation of strong eukaryotic and faint anaerobic fungal PCR bands from BM3E string, but not from BM3E leachate. The finding is consistent with Lockhart *et al*. (2006), who only reported the distribution of anaerobic fungi in landfill cells where cellulose was the primary source of carbon. Leachate in BM3F, 3G, 3H and 1J is low in organic matter but high in inorganic matter, and we

suggest this is why the leachate samples were tested to be positive for the presence of eukaryotes using PCR. It has been suggested previously that conditions such as low organic matter coupled with high inorganic matter allow for the proliferation of all eukaryotes, especially those with larger cells, as they tend to be better competitors (Jacquet *et al*., 2002). These eukaryotes might include anaerobic fungi and ciliated protozoa in such an environment. Since microbial diversity in landfill is very poorly characterised, studies describing protozoa in landfill are rare.

Table 3.3. Table showing the landfill leachate chemical data from all risers across the landfill sites (obtained from Merseyside Waste Disposal Authority and Biffa Waste Services Limited), and the current groundwater trigger levels for the compounds listed, except for biochemical oxygen demand (BOD) and chemical oxygen demand (COD).

	BD 3/4	**BM3E**	**BM3F**	**BM3G**	**BM3H**	**BM1J**	**Groundwater**
BOD	349	260	5	4	4	20	-
COD	2650	1000	75	190	94	-	-
Iron	0.51	2180	300	1580	450	610	0.2
Sodium	2410	140	350	1200	230	620	170
Potassium	706	12	60	250	110	310	12
Magnesium	102	63	93	120	110	130	50
Chloride	3290	420	400	2000	300	670	250

All measurements in mgL^{-1}
- No data

3.3 Culturing anaerobic fungi from landfill leachate

A modified version of the Hungate technique (Hungate, 1950) was developed and an attempt was made to culture and isolate representatives belonging to the strictly anaerobic cellulolytic fungal order *Neocallimastigales*. The method was developed for the isolation of anaerobic mesophilic cellulolytic bacteria, including Clostridia, and as such a few modifications were necessary to render it appropriate for the culturing of strict anaerobes that are particularly sensitive to oxygen exposure.

3.3.1 Initial testing and enrichment set-up

Due to the obligate anaerobic lifestyle of *Neocallimastigales* fungi, it was critical to ensure that the culture medium was completely reduced. Different combinations of reducing agents and the effect of boiling the media prior to incubation were investigated to determine which condition produced a stable reduced environment. The mock medium was prepared by the addition of 0.0001% resazurin (w/v) final concentration to 20 ml ddH_2O per culture tube. Resazurin dye was used as a redox indicator, as it turns from dark blue to colourless when the conditions change from aerobic to anaerobic (Hungate, 1950). Combinations of reducing agents reported in culturing experiments, namely cysteine hydrochloride, sodium pyruvate and sodium thioglycolate (Hungate, 1950; Lowe *et al*., 1985; Lowe *et al*., 1987), were tested. The media were either boiled for 10-15 minutes or not boiled at all. The samples were tested in 22 ml glass culture tubes (Supelco), with rubber septa (Supelco) between the tubes and the screw-on lids to ensure that the cultures remained airtight during incubation. Anaerobic gas mix (5% CO_2, 10% H_2 and 85% N_2, BOC gases) was bubbled through the mock media and was also used to fill the head space of the culture tubes. The tubes

were set up in triplicate for each treatment and were incubated overnight at room temperature, prior to being autoclaved.

It was determined that pre-boiling the mock media maintained anaerobic conditions for longer in all the culture tubes. Some of the reducing agent treatments tested only produced anaerobic conditions for 5 days (sodium thioglycolate only), 8 days (sodium thioglycolate and cysteine hydrochloride) and 12 days (sodium thioglycolate and sodium pyruvate), whereas some treatments did not create anaerobic conditions at all (sodium pyruvate only; cysteine hydrochloride only; cysteine hydrochloride and sodium pyruvate). The reducing agent combination of cysteine hydrochloride (0.005% w/v), sodium pyruvate (0.01% w/v) and sodium thioglycolate (0.01% w/v) produced anaerobic conditions that lasted for 2 weeks after anaerobic gas was bubbled through the media. When the media were pre-boiled and autoclaved after the addition of reducing agents, the treatment with all 3 reducing agents produced anaerobic conditions within the culture tubes for over 5 weeks, which was better than any other treatment. As such, this reducing agent combination was chosen for the culturing experiment.

Despite filling the headspace of the culture tubes with anaerobic gas mix, appearance of a pink colour within the tubes after 5 weeks suggested that the conditions had slowly reverted to oxic. Since incubation of up to 12 weeks was required for the culturing, an anaerobic cabinet (Don Whitley, Fig 3.1) was chosen for incubation of the culture tubes to ensure prolonged anaerobiosis. The cabinet was filled with the same anaerobic gas mix that was bubbled through the media and was used to fill the headspace in the culture tubes. Moreover, subculturing within the cabinet would prevent introduction of any trace amount of oxygen into the culture tubes. An antibiotic solution consisting of 0.007% w/v each of chloramphenicol, streptomycin and ampicillin was used to

discourage extensive bacterial growth and allow for the slow growing fungi to proliferate (Lowe *et al*., 1985; Kurakov *et al*., 2011). Six different cellulosic substrates (0.4 g each) were selected: powdered cellulose (Avicel), soluble carboxymethylcellulose (CMC), the cellulose repeating unit cellobiose, dewaxed cotton string, thin strips of pure cellulose filter paper (Whatman) and fibrous lignocellulosic hay.

Since a weak nested fungal PCR positive was detected, leachate collected from riser BM1J was chosen as the inoculum mixed with a highly reduced culture medium, incorporating a reducing agent combination of cysteine hydrochloride (0.005% w/v), sodium pyruvate (0.01% w/v) and sodium thioglycolate (0.01% w/v), and the experiment was performed in 20 ml airtight tubes that were filled with anaerobic gas mix, as described in section 2.4. Two rounds of subculturing at an interval of 6 weeks each followed the initial culture. PCR checks were performed on each round of culture after 8 weeks of incubation to detect the presence of eukaryotes, *Neocallimastigales*, bacteria, *Fibrobacter* and *Clostridium* cluster III.

Figure 3.1. The Don Whitley anaerobic cabinet used to incubate the reduced culture tubes for a period of up to 12 weeks, connected to cylinders of anaerobic gas mix (BOC gases).

3.3.2 PCR analysis

Following 8 weeks of incubation in the anaerobic cabinet, 5 ml of culture medium was recovered from one replicate of each initial culture tube, and was subsequently centrifuged at 10,000 *g* for 5 minutes to concentrate the biomass. The concentrated biomass was then used to extract DNA using the Griffiths *et al*. (2000) method, as described in section 2.2.2. However, DNA yields were determined to be low, with only between 30-100 ng of DNA extracted from the various cultures and no DNA extracted from the culture with powdered Avicel substrate. In order to extract enough DNA to perform PCR amplification, the contents of one whole replicate culture tube from each treatment were used for DNA extraction. This allowed for 20 ml of culture medium as well as any solid substrate within the tubes to be subjected to DNA extraction. The DNA yield was still concluded to be low, with the lowest yield of 50 ng from Avicel substrate and highest yield of 300 ng from hay substrate, but was sufficient to perform PCR reactions. Direct and nested PCR amplification was performed to detect eukaryotes, *Neocallimastigales*, bacteria, *Fibrobacter* and *Clostridium* cluster III, and the results are presented in Table 3.4.

In the initial culture, bacterial DNA was detected in 4 out of the 6 substrates (Avicel, cellobiose, dewaxed cotton and filter paper) as well as the no substrate with antibiotics and no substrate without antibiotics controls using universal bacterial 16S rRNA gene amplification. *Fibrobacter* spp. could not be detected from any culture using direct PCR, and could only be amplified from the no substrate without antibiotics control using nested PCR. Direct amplification of *Clostridium* cluster III only produced a positive PCR reaction from the no substrate without antibiotics control, whereas the nested PCR produced a positive reaction from all cultures except from the hay substrate. However, eukaryotes were detected in only the no substrate without antibiotics control

and no *Chytridiomycetes* were detected in any of the cultures using either direct or nested PCR. It is worth mentioning that the *Chytridiomycete* PCR primer set had previously been validated by the successful amplification of anaerobic fungi from a sheep rumen sample.

The first subculture was set up from the initial culture following 6 weeks of incubation, and the contents of one whole culture tube from each treatment were again used for DNA extraction. The DNA yield was determined to be very low, with between only 30 and 80 ng of total DNA obtained from each of the culture tubes. PCR analysis was performed again in the same manner as for the initial culture and the data is presented in Table 3.5.

Bacteria were detected only in culture tubes with cellobiose, filter paper and string substrates, while *Fibrobacter* and *Clostridium* cluster III could not be detected using direct PCR. *Fibrobacter* could not be detected using nested PCR either and PCR bands from *Clostridia* were only observed in 4 out of 6 substrates using nested PCR (Avicel, cellobiose, filter paper and dewaxed string). No eukaryotes or *Chytridiomycetes* were detected in any of the substrates using either direct or nested PCR. It is worth noting that not enough DNA could be extracted from the CMC culture tube or either of the no substrate controls to be able to perform PCR analysis, and hence that data could not be included here. This problem persisted with the second subculture, as no meaningful DNA could be extracted from any of the culture tubes, resulting in no PCR data from this round of culturing. It was decided that further subculturing was not required.

Table 3.4. Table showing the results of the PCR analysis performed on the initial culture tubes from the culturing experiment. DNA was extracted from one replicate tube each of the six substrates, as well as the relevant controls. Direct PCR was performed unless otherwise stated.

	Avicel	**CMC**	**Cellobiose**	**Dewaxed cotton**	**Filter paper**	**Hay**	**No substrate + antibiotics**	**No substrate no antibiotics**
General bacteria	+	-	+	+	+	-	+	+
General eukaryotes	-	-	-	-	-	-	-	+
Fibrobacter	-	-	-	-	-	-	-	-
Fibrobacter (nested)	-	-	-	-	-	-	-	+
***Clostridium* cluster III**	-	-	-	-	-	-	-	+
***Clostridium* cluster III (nested)**	+	+	+	+	+	-	+	+
Anaerobic fungi	-	-	-	-	-	-	-	-
Anaerobic fungi (nested)	-	-	-	-	-	-	-	-

+ Presence of PCR product
- Absence of PCR product

Table 3.5. Table showing the results of the PCR analysis performed on the first subculture tubes from the culturing experiment. DNA was extracted from one replicate tube each of the six substrates, as well as the relevant controls. Direct PCR was performed unless otherwise stated.

	Avicel	**CMC**	**Cellobiose**	**Dewaxed cotton**	**Filter paper**	**Hay**	**No substrate + antibiotics**	**No substrate no antibiotics**
General bacteria	-	ND	+	+	+	-	ND	ND
General eukaryotes	-	ND	-	-	-	-	ND	ND
Fibrobacter	-	ND	-	-	-	-	ND	ND
Fibrobacter (nested)	-	ND	-	-	-	-	ND	ND
***Clostridium* cluster III**	-	ND	-	-	-	-	ND	ND
***Clostridium* cluster III (nested)**	+	ND	+	+	+	-	ND	ND
Anaerobic fungi	-	ND	-	-	-	-	ND	ND
Anaerobic fungi (nested)	-	ND	-	-	-	-	ND	ND

+ Presence of PCR product
- Absence of PCR product
ND No data

The inability to successfully amplify *Clostridium* cluster III DNA from any culture apart from the no antibiotic control is significant, as it suggests that the antibiotic cocktail added to the culture medium was at least partially successful in lowering the prokaryotic abundance. Previous PCR results obtained from this lab have consistently shown clostridia to be abundant enough in the leachate to be detected using direct PCR alone (Van Dyke & McCarthy, 2002; Lockhart *et al*., 2006; McDonald *et al*., 2008; McDonald *et al*., 2012).

PCR analysis of the initial culture seems to imply that the reduced medium used is suitable for culturing of anaerobic eukaryotes, as eukaryotes were detected in the no substrate without antibiotics control. However, the absence of any detectable eukaryotes in the no substrate with antibiotics control suggests that prokaryotes are potentially required to co-exist in the culture medium, as the complex interactions between different groups of bacteria and eukaryotes might be essential to maintain the metabolic balance in a heterogeneous environmental microbial community, allowing eukaryotes to proliferate (Jacquet *et al*., 2002).

Failure to extract any DNA from no substrate culture tubes from the first subculture suggests that a source of cellulose might be required for the continued metabolism of the microbial community within the tubes, since only 2 ml of the initial culture was added to fresh reduced culture medium. Although little, DNA was successfully extracted from some first subculture tubes where an external cellulose source was added. For the purpose of future experiments, it is suggested that culturing is performed in 60 ml serum bottles (Supelco) with crimp seals, as the use of 50 ml culture media will facilitate increased microbial biomass and hence provide better DNA yield to perform robust PCR analysis on. It is worth noting that the anaerobic gas mix used in this study is slightly less dense than air and could potentially allow oxygen

contamination, although the use of a strong cocktail of reducing agents as well as the anaerobic cabinet allows for any excess oxygen present in the tubes to be mopped up. CO_2 can be used as an alternative gas to maintain anoxia within the tubes in future experiments, albeit with a bicarbonate-based buffer in the reducing medium (Hungate, 1950).

3.4 Determining the composition of the amplified eukaryotic community in landfill leachate

The inability to successfully amplify any *Neocallimastigales* DNA from the culture tubes was surprising, as eukaryotic PCR products were observed from both the incubated string and the leachate from some Bidston Moss samples. In order to determine the taxonomic identity of the amplified eukaryotic PCR products, positive PCR products obtained each from the string incubated in leachate 3E (3ES), the string incubated in leachate 1J (1JS) and the 1J leachate (1JL) itself were used for molecular cloning.

3.4.1 Molecular cloning, sequencing and bioinformatic analysis

Molecular cloning of the amplified rDNA PCR amplification products was performed as described in section 2.10, using the pGEM-T Easy plasmid (Promega) as vector and the JM109 high efficiency chemically competent cells as host. Blue-white screening was performed to determine the identity of the clones carrying the insert, and 6 white colonies each from the 3 samples were chosen at random to extract the plasmid DNA from. Restriction digestion was performed using EcoRI restriction enzyme to ensure that the insert in each plasmid was of the correct size (~1,600 bp). Isolated plasmids were then sent to GATC Biotech, Germany for single read sequencing to

obtain some preliminary data before further inserts would be sequenced. The 18 sequences were aligned against the NCBI nr nucleotide database using BLAST (Altschul *et al*., 1990) with an E-value cut-off of 0.001, and the results are presented in Table 3.6.

Table 3.6. Table detailing the results of the BLAST search performed to align the eighteen cloned 18S rRNA gene amplicons, six each sequenced from the three clone libraries, against the NCBI nr database.

Sample	Sequence number	Closest BLAST hit
Carboy 3E, string	1	*Metopus palaeformis*
	2	*Metopus palaeformis*
	3	Uncultured marine picoeukaryote/ *Trimastix pyriformis*
	4	*Trimastix pyriformis*/ uncultured Chytridiomycota clone
	5	*Demodex brevis*
	6	Uncultured fungal clone
Carboy 1J, leachate	1	*Metopus palaeformis*
	2	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*
	3	*Metopus palaeformis*
	4	*Metopus palaeformis*
	5	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*
	6	*Metopus palaeformis*
Carboy 1J, string	1	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*
	2	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*
	3	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*/ Brachonella sp.
	4	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*/ Brachonella sp.
	5	*Metopus palaeformis*
	6	*Nyctotherus* sp./ *Clevelandella* sp./ *Metopus palaeformis*/ Brachonella sp.

Of the 18 inserts sequenced, 7 had the protozoan *Metopus palaeformis* as the closest BLAST hit match, with a further 7 having a closest BLAST hit match belonging to the protozoan *Nyctotherus* sp. Only 1 insert was a possible anaerobic fungal hit, suggesting that the PCR amplified eukaryotic community in the Bidston Moss leachate samples 3E and 1J is dominated by anaerobic protozoa. This is particularly interesting as literature focussing on the biodiversity or function of anaerobic protozoa in municipal landfill leachate is practically non-existent.

Metopus palaeformis are anaerobic, free-living ciliated protozoa that have endosymbiotic methanogenic bacteria distributed throughout their cytoplasm (Lynn, 2010). They have previously been found in anaerobic waste treatment systems, marine lagoons and sulphide-rich lakes (Lynn, 2010). Metopids form a monophyletic clade with the Nyctotherids, and both belong to the ciliate order of Armophorida (Lynn & Wright, 2013). *Nyctotherus* are also anaerobic ciliated protozoa that have previously been found to be either free-living or as commensals in the intestinal tract of termites, wood-feeding cockroaches and frogs (Lynn, 2010). Their role in cellulose degradation in the intestines of wood-feeding cockroaches has been documented by Gijzen *et al.* (1994). It is also worth noting that 5 out of 6 inserts sequenced from sample 1JS had *Nyctotherus* as the closest BLAST hit match, whereas only 2 out of 6 inserts from sample 1JL had *Nyctotherus* as the closest BLAST hit match, and these ciliates have been reported to be actively involved in lignocellulosic biomass.

3.4.2 Analyses of 454 pyrosequenced datasets

Following the sequencing and bioinformatic analysis of a small subset of clones from the above mentioned clone libraries, a further ~50 clones each from the 3 clone libraries could be investigated. However, it was decided that the analysis of a 454

pyrosequenced dataset would be better in elucidating the taxonomic make-up of the eukaryotic fraction of the landfill microbial community responsible for cellulose degradation, as an output of thousands of reads would provide better resolution.

3.4.2.1 Preparation, sequencing and bioinformatic analyses of the datasets

Landfill leachate collected from Bromborough Dock as well as dewaxed cotton string incubated in the leachate were used for extraction of total community DNA, followed by PCR amplification using specific primers targeting the internal transcribed spacer (ITS) region of the 18S rRNA gene in eukaryotes. Multiplex identifiers (MIDs) were added to the resultant 410 bp long amplicons, which were subsequently mixed with other amplicons generated from bacterial and archaeal DNA from landfill leachate and landfill cotton. The amplicons were sequenced using the Roche GS FLX Titanium, with a ¼ plate each loaded with all amplicons generated either from landfill leachate or cotton incubated in landfill leachate. DNA extraction and amplicon library preparation was performed by Dr David Rooks and the samples were sequenced at the Centre for Genomic Research (CGR) in the University of Liverpool, where the bioinformatic analyses were performed by Dr Xuan Liu using Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso *et al*., 2010).

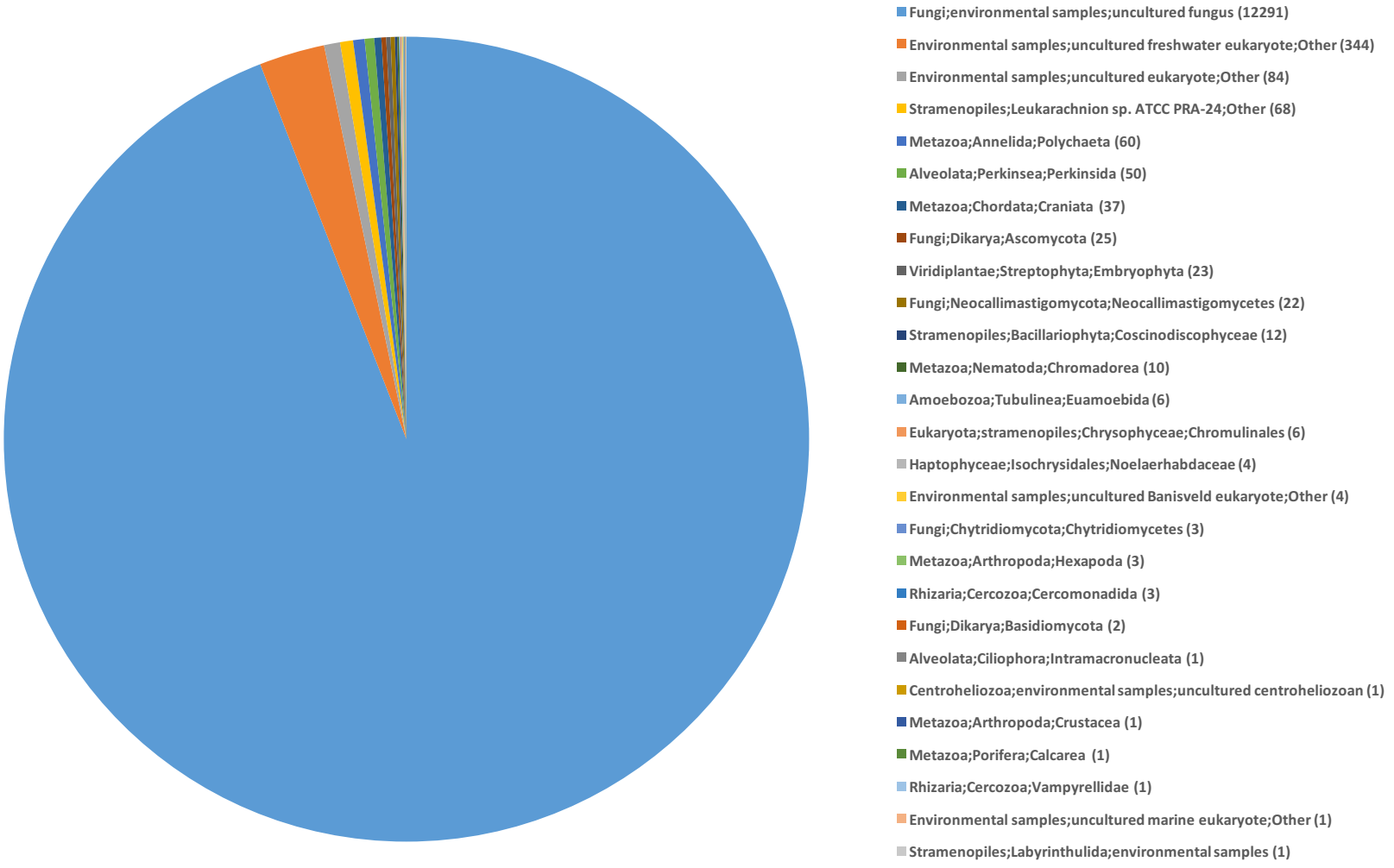
The sequencing output was denoised and was subsequently subjected to rigorous quality control, including removal of sequences with: incorrect length, number of ambiguous bases exceeding 6, mean quality score below 25, homopolymer runs exceeding 6 bases, and mismatches in the primer sequence. The sequence libraries were then split based on the MID sequence in order to separate the 18S rRNA amplicons from bacterial and archaeal amplicons. A total of 12,704 and 13,064 18S rRNA sequences were eventually left for Operational Taxonomic Unit (OTU) picking against

the SILVA database (Pruesse *et al*., 2007), from landfill leachate and cotton incubated in landfill leachate, respectively.

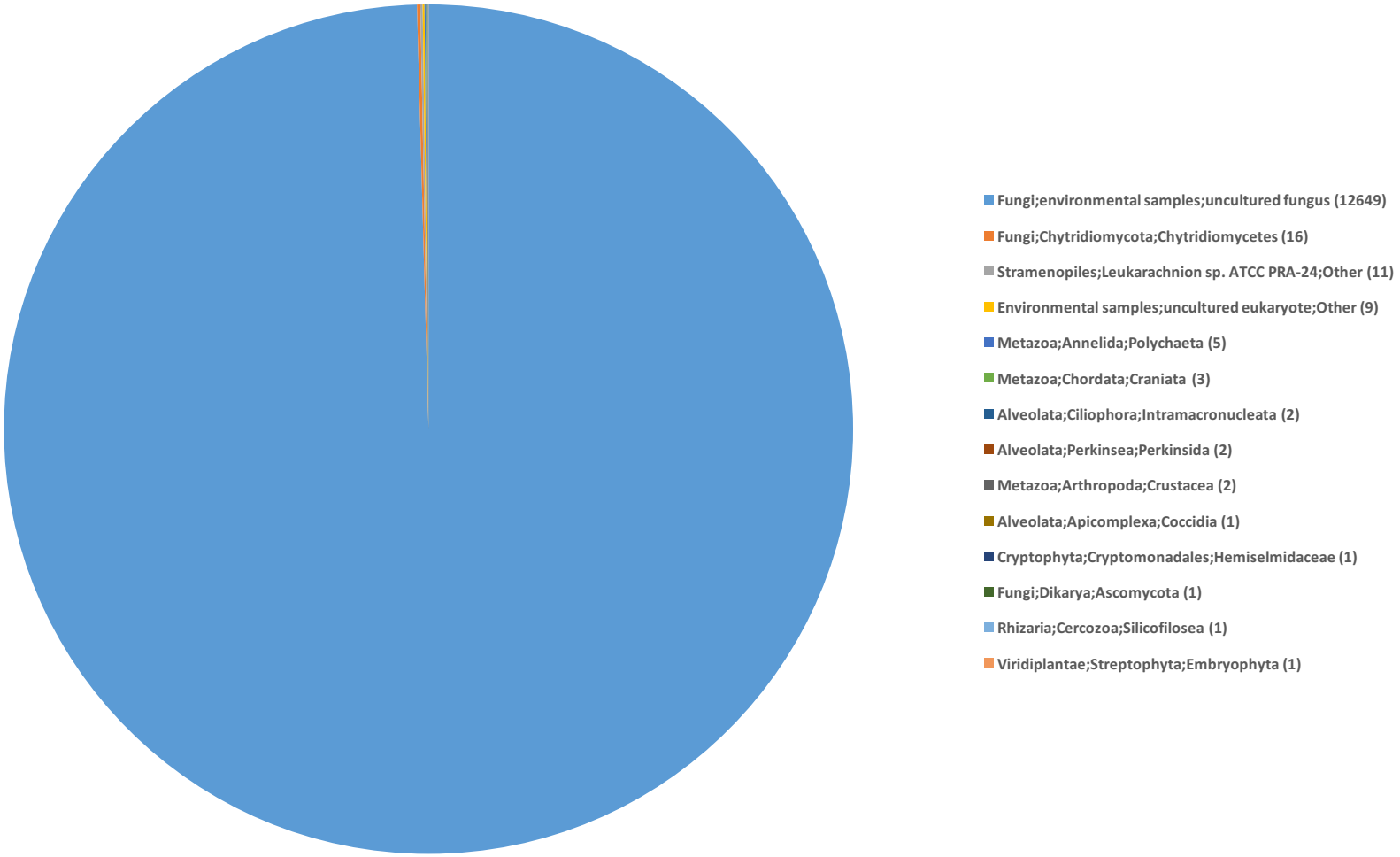
3.4.2.2 Results

Taxonomic classification of the eukaryotic component of the microbial community present in Bromborough Dock leachate, as well as in the biofilm colonising the dewaxed cotton incubated in the leachate is presented in Fig. 3.2. Figure 3.2 (A) suggests that fungi are the most abundant eukaryotes associated with the cotton incubated in the leachate, making up 94.1% of the total biodiversity. Uncultured freshwater eukaryotes are the only other constituent making up more than 1% of the eukaryotic biodiversity (2.6%). Sequences corresponding to *Neocallimastigales* fungi account for 0.2% of the total diversity (22 sequences), and 3 sequences had *Chytrodiomycetes* fungi as the closest match in the database. The singular sequence belonging to ciliated protozoa from the phylum Ciliophora has the closest hit corresponding to Peniculids, an order consisting of heterotrophic protists such as *Paramecium* (Lynn, 2010).

Fungi also dominate the eukaryotic community populating the landfill leachate in Bromborough Dock, as they account for 99.6% of total eukaryotes present (Figure 3.2 (B)). In contrast to the colonised cotton dataset, no sequences were determined to have a closest match directly corresponding to *Neocallimastigales* fungi, perhaps further illustrating the important role these microbes play in colonisation and subsequent decomposition of lignocellulosic biomass. However, *Chytrodiomycetes* fungi account for 0.1% of the total eukaryotic diversity in leachate (16 sequences). The two ciliated protozoa hits in the dataset correspond to microbes in subclass Haptoria, which comprise primarily of heterotrophic protozoa (Lynn, 2010).


Fungi;environmental samples;uncultured fungus (12291)
Environmental samples;uncultured freshwater eukaryote;Other (344)
Environmental samples;uncultured eukaryote;Other (84)
Stramenopiles;Leukarachnion sp. ATCC PRA-24;Other (68)
Metazoa;Annelida;Polychaeta (60)
Alveolata;Perkinsea;Perkinsida (50)
Metazoa;Chordata;Craniata (37)
Fungi;Dikarya;Ascomycota (25)
Viridiplantae;Streptophyta;Embryophyta (23)
Fungi;Neocallimastigomycota;Neocallimastigomycetes (22)
Stramenopiles;Bacillariophyta;Coscinodiscophyceae (12)
Metazoa;Nematoda;Chromadorea (10)
Amoebozoa;Tubulinea;Euamoebida (6)
Eukaryota;stramenopiles;Chrysophyceae;Chromulinales (6)
Haptophyceae;Isochrysidales;Noelaerhabdaceae (4)
Environmental samples;uncultured Banisveld eukaryote;Other (4)
Fungi;Chytridiomycota;Chytridiomycetes (3)
Metazoa;Arthropoda;Hexapoda (3)
Rhizaria;Cercozoa;Cercomonadida (3)
Fungi;Dikarya;Basidiomycota (2)
Alveolata;Ciliophora;Intramacronucleata (1)
Centroheliozoa;environmental samples;uncultured centroheliozoan (1)
Metazoa;Arthropoda;Crustacea (1)
Metazoa;Porifera;Calcarea (1)
Rhizaria;Cercozoa;Vampyrellidae (1)
Environmental samples;uncultured marine eukaryote;Other (1)
Stramenopiles;Labyrinthulida;environmental samples (1)

A



B

Figure 3.2 Taxonomic classification of eukaryotes colonising (A) cotton incubated in BD leachate, as well as (B) BD leachate itself, using PCR-generated amplicons of the 18S rRNA gene. Sequenced using the Roche GS FLX Titanium and analysed using QIIME (Caporaso *et al*., 2010).

The results are surprising, as ciliated protozoa that made up majority of the initial clone library data were determined to be absent in both the leachate as well as the dewaxed cotton. Furthermore, the results here contradict the results from the PCR analysis performed on BD leachate collected from riser 3/4 where no positive reactions were observed for the presence of anaerobic fungi, a result that is also consistent with previous data generated in our lab (McDonald *et al*., 2012).

However, it is possible that the leachate sample used to generate amplicons sequenced using the Roche GS FLX Titanium was collected from riser 2 at Bromborough Dock, as it was the only sample that yielded a positive PCR product for anaerobic fungi (albeit through nested PCR). The more regular detection of anaerobic cellulolytic fungi from Bidston Moss landfill leachate can be attributed to an elevated level of biodegradable cellulose biomass, as paper mill waste has been frequently used as cover for the landfill site (McDonald *et al*., 2012). This further reiterates the irregular nature of anaerobic fungi distribution in landfill, and that a greater than average load of cellulosic biomass is perhaps required to facilitate their growth and subsequent detection using molecular biological techniques such as PCR.

It is interesting to note that the overwhelming majority of the sequences from both samples return uncultured representatives as the closest match in the database, pointing not only to the fact that analysis of landfill community composition remains understudied, but also to the lack of cultured representatives belonging to anaerobic fungi, many of which could play environmentally significant roles in cellulosic biomass degradation.

3.5 Discussion

The taxonomic identity of eukaryotes colonising the string incubated in Bidston Moss leachate samples 3E and 1J, as well as the leachate 1J itself was investigated. Initial clone library data had suggested that ciliated protozoa of class Armophorea constituted the major proportion of eukaryotes present in such an environment, as 14 out of 18 sequences returned *Metopus* sp. or the phylogenetically close relative *Nyctotherus* sp. as the nearest match using Blast analysis. *Metopus* spp. have previously been found in anaerobic waste treatment systems (Lynn, 2010) while *Nyctotherus* spp. have been reported to be involved in cellulose degradation in the intestines of wood-feeding cockroaches (Gijzen *et al*., 1994), yet no literature exists on their potential role in lignocellulose decomposition in a landfill environment. While ciliated protozoa have been documented to perform cellulose degradation in the rumen, those are classified under a different class altogether; Litostomatea and not Armophorea (Wright *et al*., 1997).

It is possible that members of class Armophorea are present in other anoxic environments also and design of specific primers could allow for PCR-mediated analysis of their environmental distribution, potentially allowing for elucidation of their role in cellulose degradation. Efforts could also be made to culture and isolate representatives of ciliated protozoa present in landfill leachate with a view to high throughput genome sequencing, as this could allow for discovery of potentially novel glycoside hydrolases. Narayanan *et al*. (2007) reported the isolation of *Metopus* sp. from an anaerobic reactor and maintenance of a pure culture using ciliate mineral medium supplemented with 1% wheat powder suspension.

Culturing of anaerobic fungi is technically challenging, particularly given the need for frequent sub-culturing of working cultures to ensure viability. Viability has

been observed to vary from strain to strain, with some reports of cultures retaining viability for ~15 days in media containing particulate substrates such as reed canary grass (Haitjema *et al*., 2014). Others have suggested that regular sub-culturing, as frequent as once every 2 days, is the key to maintaining viability of cultures in highly reduced medium (personal communication with M. Cougar, co-author on Youssef *et al*., 2013). By comparison, a vastly different sub-culturing time of 8 weeks was used in this study. This could be the most likely explanation for the inability to isolate any cultured anaerobic cellulolytic fungi representatives in this study.

Despite the use of highly reduced culturing medium to supplement the landfill leachate, it is almost impossible to mimic the exact environmental conditions required for the growth and proliferation of these extremely fastidious microbes. Furthermore, anaerobic fungi are known to be highly sensitive to the presence of oxygen. Periodic sample collection from the carboys containing the landfill leachate and the incubated string could introduce enough oxygen to cause the anaerobic fungi to sporulate as well as potentially settle at the bottom of the carboy. The fungi might even have been outcompeted in the collected leachate over time, causing them to perish altogether from the sample. These factors might also explain why weak nested PCR positive results for anaerobic fungi could not be replicated 2 years later from the BM1J sample. Moreover, sporadic distribution of anaerobic cellulolytic fungi has been observed in landfill sites and these microbes were determined to not be present in the majority of the BD samples, as has been suggested by McDonald *et al*. (2012).

The optimal time to extract high quality DNA from growing monocentric and polycentric fungi is between 2-3 days (Tsai & Calza, 1992). This is because nucleic acids are concentrated in zoospores, whilst the remaining nucleic acids in the rhizomycelium are prone to contamination from polysaccharides. DNA isolation earlier

than this time will result in very low yield, whereas those after will result in degraded DNA due to inherent instabilities associated with low GC genomes (Brownlee, 1989; Nicholson *et al*., 2005). Further contamination of the DNA from polysaccharides can occur during chemical extraction methods, owing to the physiology and typical growth conditions of these anaerobic fungi (Brownlee, 1988; Brownlee, 1989). These details offer an explanation as to why molecular techniques such as PCR might have failed to provide comprehensive evidence if the fungi were indeed present in some of the cultures.

Studies identifying anaerobic fungal cellulases and seeking to express them in heterologous hosts are emerging rapidly (Chen *et al*., 2012a; Comlekcioglu *et al*., 2010; O'Malley *et al*., 2012). However, many of these enzymes under investigation demonstrated low or undetectable activity (Harhangi *et al*., 2003; O'Malley *et al*., 2012), potentially due to the lack of post-translational machinery required for the activity of these enzymes in prokaryotic hosts. Moreover, protocols for stable genetic manipulation of these microbes are practically redundant (Durand *et al*., 1997), making the introduction of gene assemblies improbable in the near future. This makes genomic and transcriptomic analyses the most likely methods for unravelling the considerable bioprocessing potential of anaerobic fungi. However, molecular analysis is arduous due to the fact that their genomes have the highest AT content of any organism identified to date (Brownlee, 1989; Youssef *et al*., 2013). Yet, recent technological advances in high throughput sequencing should enable us to not only assign biochemical function, but also to reconstruct metabolic networks in the future.

Analysis of the recently sequenced genome of *Orpinomyces* reiterates the fantastic capability of these anaerobic microbes for cellulose hydrolysis (Youssef *et al*., 2013). It is essential that attempts to culture and isolate cellulolytic fungi are continued

from understudied anoxic environments, as it will lead to the genome sequencing of further representatives of the most potent cellulose degrading microorganisms on earth. This is particularly crucial as a large proportion of *in silico* data obtained from next generation sequencing of metagenomic samples increasingly comprises of sequences with no match in databases, including sequences corresponding to uncultured microbes or genes encoding hypothetical proteins.

Chapter 4

Metagenomic and metatranscriptomic analyses of cellulolytic microbial communities and cellulases in landfill leachate

4.1 Background

Metagenomic and metatranscriptomic approaches allow us to compare and contrast the genetic potential with the *in situ* functional activity in a complex community of microbes, and will prove to be integral tools in molecular microbial ecology studies aimed at interpreting biology at the aggregate level. However, analyses of high throughput sequencing data is challenging, and it is important to utilise the appropriate analytical tools in the correct manner in order to draw biologically pertinent conclusions. Trimming and quality filtering of sequence output from high throughput sequencing platforms is essential before performing analysis, particularly in the case of 454 pyrosequencing data, as there is a significant increase in noise along the length of the read (Balzer *et al*., 2010). PANDAseq (Masella *et al*., 2012) and Prinseq (Schmieder & Edwards, 2011) allow for filtering of reads based on a variety of parameters, including read length, number of base substitutions, artificial duplicates and entropy score.

Assembly of metagenomic data is tricky given the low coverage, relatively short read length and in most cases, the absence of a reference genome or metagenome (Logares *et al*., 2012). Genovo (Laserson *et al*., 2011) and MetaVelvet (Namiki *et al*., 2012) are dedicated *de novo* metagenome assemblers designed for use with 454 pyrosequencing data and Illumina data, respectively. However, fast and efficient alignment of millions of short metagenomic reads to reference sequences is achievable

through the use of short sequence aligners such as Bowtie (Langmead *et al*., 2009), Bowtie2 (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009). While alignment of sequences against dedicated rRNA sequence databases such as SILVA (Pruesse *et al*., 2007), Ribosomal Database Project (RDP) (Cole *et al*., 2007) and Greengenes (DeSantis *et al*., 2006) can be performed using Blast to reveal the phylogenetic make-up of a microbial community, tools such as Kraken (Wood & Salzberg, 2014) and MetaPhlAn (Segata *et al*., 2012) specialise in generating such information from metagenomic reads.

Gene prediction algorithms such as Orphelia (Hoff *et al*., 2009), MetaGeneMark (Zhu *et al*., 2010) and FragGeneScan (Rho *et al*., 2010) have proved to be highly efficient in predicting Open Reading Frames (ORFs) in unassembled shotgun metagenomic or metatranscriptomic reads. This is usually followed by querying the predicted ORFs against a protein database, such as NCBI nr, SwissProt and pfam, using either Blastx or HMMer algorithms to determine the functional profile of the microbial community. Stand-alone software such as MEGAN (Huson *et al*., 2007) and CARMA (Gerlach *et al*., 2009) and the web-based MG-RAST (Meyer *et al*., 2008) allow for both taxonomic as well as functional inferences to be made from shotgun metagenomic and metatranscriptomic data. Metabolic pathways from the whole microbial community can be reconstructed by mapping protein-coding reads against curated databases such as KEGG (Kanehisa & Goto, 2000), although successful metabolic pathway reconstruction largely depends on the length of the sequences and the depth of the sequencing run.

While metatranscriptomic overview of microbial communities exists from a number of environments, not all studies have compared them to the corresponding metagenomes. Previous data generated in our lab has demonstrated the presence of

cellulolytic microbes in landfill leachate (Van Dyke & McCarthy, 2002; Lockhart *et al*., 2006; McDonald *et al*., 2012). However, no studies have focused on the taxonomic and functional profiling of the cellulose degrading microbial community in landfill leachate to date, making the comparison between such a community's metagenome and metatranscriptome presented here unique. It was also hypothesized that deep sequencing of the mRNA of a diverse microbial community indigenous to landfill leachate could potentially lead to the discovery of novel GH-encoding genes that would ordinarily be missed by traditional sequencing approaches.

4.2 Nucleic acid extraction for high throughput sequencing

Dewaxed cotton string had been incubated in landfill leachate collected from Bromborough Dock (BD1, BD2 and BD3), and Bidston Moss (BM3E, BM3F, BM3G, BM3H and BM1J) in 10 L carboys (Nalgene). After an incubation period of 6-8 months, cotton incubated in all leachate samples except BM3H indicated clear signs of microbial biofilm formation. Furthermore, cotton incubated in leachate samples BD1, BD2, BD3, BM3E and BM1J had already been determined to be colonised by cellulolytic microbes using PCR (section 3.2). There, representatives of *Fibrobacter* spp. and *Clostridium* cluster III were detected by direct amplification of the 16S rRNA gene from all cotton samples, while only weak 18S rRNA gene amplification was observed corresponding to anaerobic fungi through nested amplification from Bidston Moss samples. As such, all cotton samples exhibiting signs of degradation were used for extraction of total community DNA and community mRNA, to be used for high throughput sequencing.

4.2.1 Extraction of metagenomic DNA

DNA & RNA were co-extracted from 0.5 g colonised cotton each incubated in leachate collected from BD1, BD2, BD3, BM3E, BM3F, BM3G and BM1J using the method described by Griffiths *et al*. (2000) (section 2.2.2). Yields of DNA ranged from 1.2 μg to 3.4 μg, and the samples were concluded to be sufficiently devoid of contaminants and impurities as the A_{260}/A_{280} values attained were greater than 1.80. The samples were pooled together after RNA removal by RNase A (section 2.2.7), and agarose gel electrophoresis was performed subsequently for visual verification of the DNA profile. A single, prominent band of DNA (>10 kb in size) was clearly visible (Fig. 4.1), suggesting the DNA was not sheared and hence, was suitable for shotgun sequencing. 1 μg of pooled metagenomic DNA was sent to the Centre for Genomic Research for sequencing, where a TruSeq DNA library was prepared before the sample was sequenced on the Illumina MiSeq platform. Paired-end sequencing was performed, with a chosen read length of 250x250 bp.

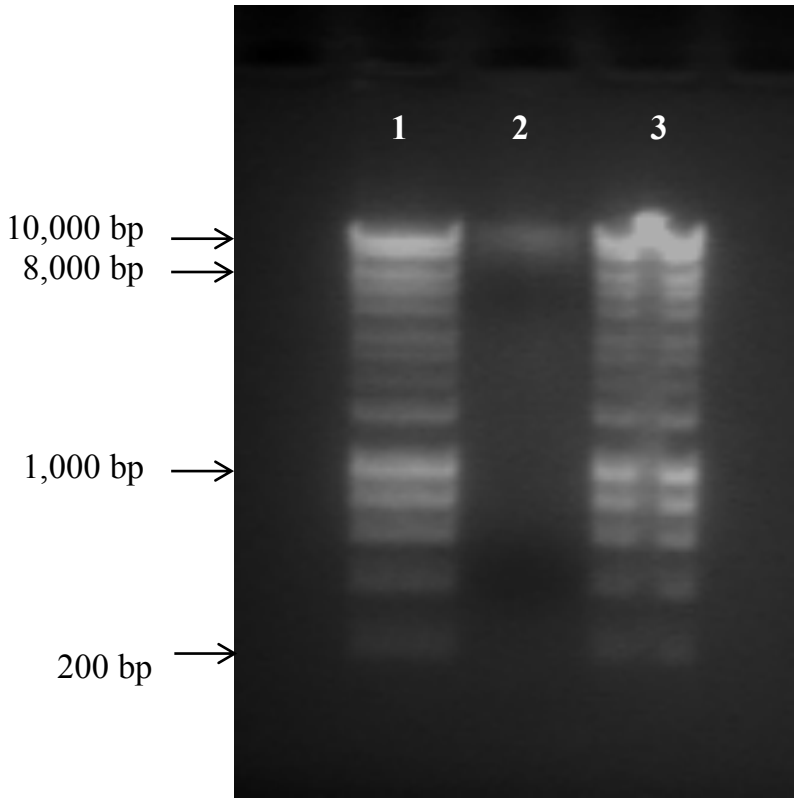


Figure 4.1 Agarose gel electrophoresis of the pooled metagenomic DNA extracted using the Griffiths *et al*. (2000) method, which was sent for sequencing on the Illumina MiSeq. Lanes 1 and 3, size marker (Hyperladder 1 kb, Bioline); lane 2, combined RNA-removed DNA extracted from cotton incubated in carboys BD1, BD2, BD3, BM3E, BM3F, BM3G and BM1J. A band of size ~ 10 kb can be seen.

4.2.2 Extraction of metatranscriptomic mRNA

Colonised cotton samples parallel to those used to generate the metagenome were chosen for the isolation of community mRNA. Total nucleic acids were extracted from 0.5 g colonised cotton each using the method described in section 2.2.8. DNA removal was performed using Turbo DNase (Invitrogen) and small RNAs, including transfer RNA (tRNA) and 5S ribosomal RNA (rRNA), were removed from the sample using the RNeasy MinElute Clean-up kit (Qiagen) following the manufacturer's instructions, as described in sections 2.2.8.1 and 2.2.8.2, respectively. Total RNA yield was determined to range between 3.5 μg and 8.5 μg, and the RNA was concluded to be sufficiently pure for downstream processing. The integrity of the extracted RNA was determined using the prokaryote total RNA 6000 (nano) assay on a 2100 Bioanalyzer (Agilent), following the manufacturer's instructions. The Bioanalyzer is a particularly useful tool for analysis of RNA degradation profile as considerably smaller amount of precious nucleic acids is required (~ 10 ng) and, as such, offers an excellent alternative to traditional agarose gel electrophoresis.

The profile of a typical RNA extract is presented in Figure 4.2. Large peaks formed of products smaller in size than the 16S rRNA subunit were clearly visible to the left of 16S rRNA subunit peak, suggesting that the RNA was degraded during the extraction protocol. The small peak farthest to the left is a marker. In order to elucidate the cause of this degradation which persisted across all samples, total RNA was extracted and prepared for analysis using the same method as described above from a pure culture of *E. coli* K12 strain to act as a control. The Bioanalyzer trace presented in Figure 4.3 showed a distinct lack of low molecular weight products, suggesting that the *E. coli* RNA extracted was of high integrity.

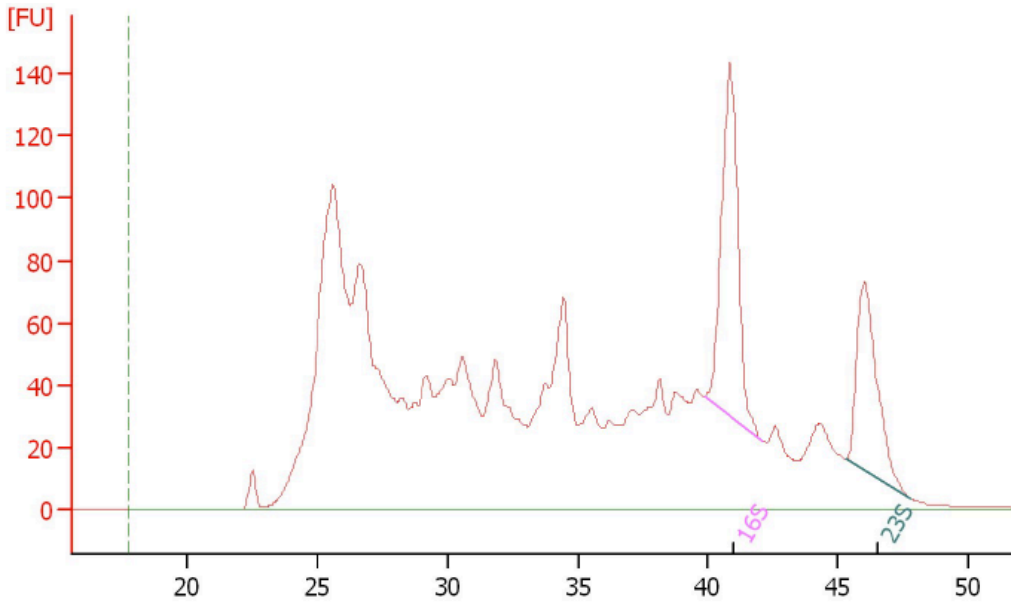


Figure 4.2 Bioanalyzer trace of a typical early total RNA extract. Multiple peaks equating to large amount of degradation products can be seen to the left of the 16S rRNA peak.

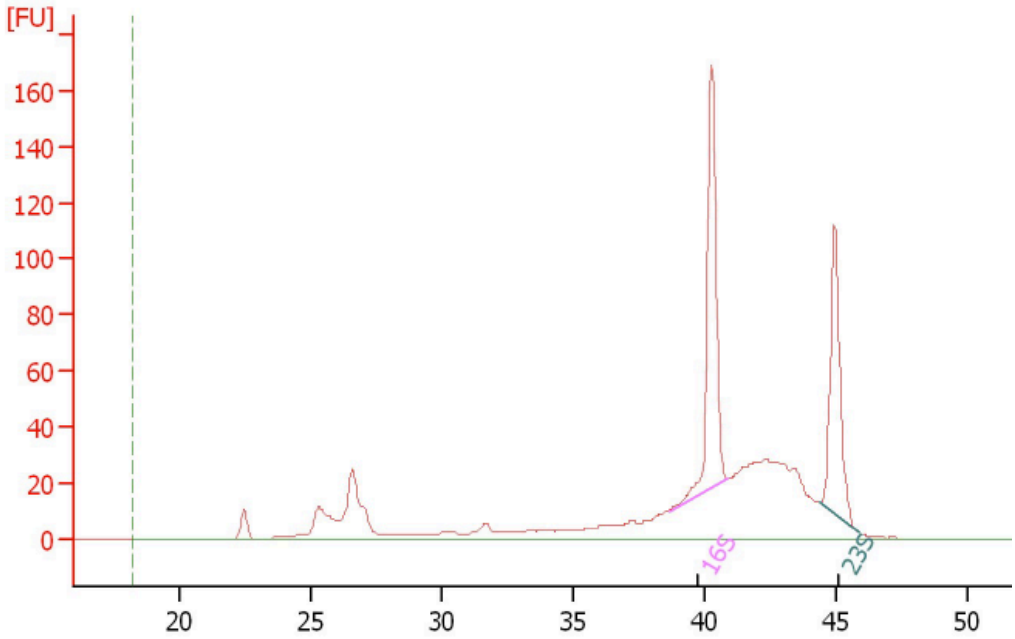


Figure 4.3 Bioanalyzer trace of the total RNA extracted from the *E. coli* K12 control strain. Very little degradation is observed to the left of prominent 16S and 23S rRNA peaks.

The hypothesis is that the incubated cotton samples had a lot of inherent RNase activity, which is not entirely surprising considering that leachate consists of a large and heterogeneous microbial community, as well as a wide range of contaminants. It was possible that bead-beating lysis of the cells embedded in the biofilm on the string released the RNA to come in direct contact with the environmental RNases associated with the sample. The extraction protocol was modified to incorporate 100 µl of 14.3 M 2-mercaptoethanol (Sigma-Aldrich) before bead-beating, allowing this strong RNase inhibitor to reduce disulphide linkages present in the RNases in the leachate. Furthermore, RNasin Plus RNase inhibitor (Promega) was added to the nucleic acids before the overnight precipitation step and also after resuspension into RNase-free water. Their combined action was deemed adequate to nullify the degradative effect of RNases at crucial steps during the extraction procedure. Bioanalysis of the RNA extracted consequently confirmed that issues with RNase-mediated RNA degradation had indeed been resolved, as a flat base-line can be seen along with the 16S and 23S rRNA subunits (Fig. 4.4).

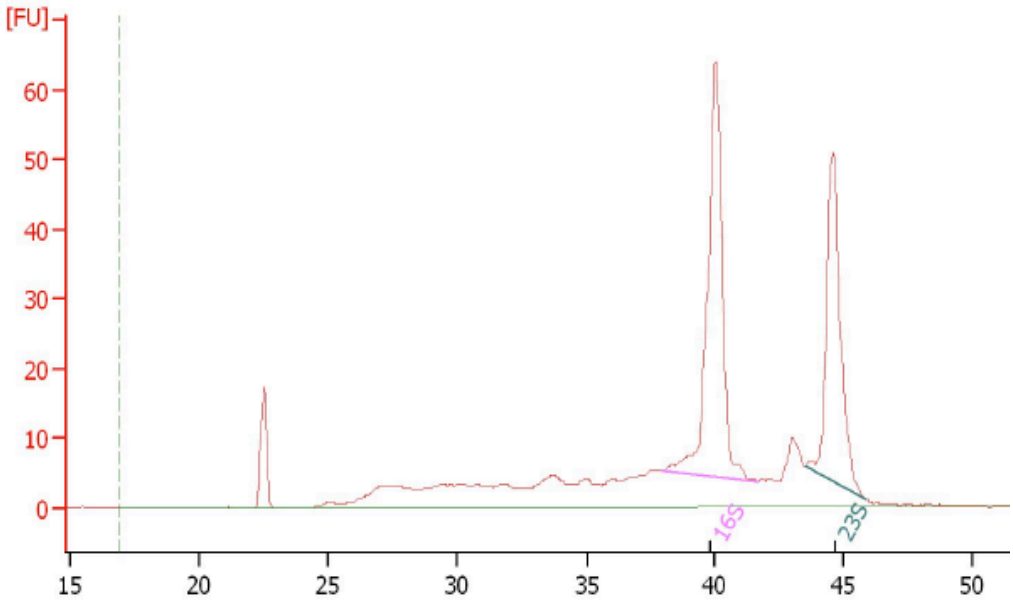


Figure 4.4 Bioanalyzer trace of total RNA extracted from string incubated in BM1J leachate (as an example) following the incorporation of 2-mercaptoethanol (Sigma Aldrich) and RNasin Plus RNase inhibitor (Promega). A flat base line suggests a lack of low molecular weight degradation products.

Since mRNA has been reported to account for only up to ~10% of total microbial RNA, rRNA depletion was incorporated into the methodology for mRNA generation to ensure that maximum sequencing depth could be channelled towards sequencing of protein-coding transcripts. rRNA depletion was performed as a two-step process, using the MICROB*Express* Bacterial mRNA Enrichment kit (Ambion) and the Terminator 5'-Phosphate-Dependent Exonuclease (Epicentre), following manufacturers' instructions. Given the somewhat fragile nature of RNA, samples were subjected to bioanalysis following each depletion protocol; this also allowed for qualitative assessment of SSU and LSU rRNA removal from the samples. The Bioanalyzer traces are presented in Figures 4.5 and 4.6, and indicate almost complete removal of small RNAs and rRNA subunits without the formation of degradation products along the way.

mRNA extracted from colonised cotton incubated in carboys BD1, BD2, BD3, BM3E, BM3F, BM3G and BM1J was pooled together following separate DNA, small RNA and SSU rRNA removal, and a profile of the sample sent to the Centre for Genomic Research for sequencing is provided in Figure 4.7. A ScriptSeq cDNA library was prepared from 250 ng of mRNA before paired-end sequencing of the sample was performed on the Illumina MiSeq platform, with a chosen read length of 250x250 bp.

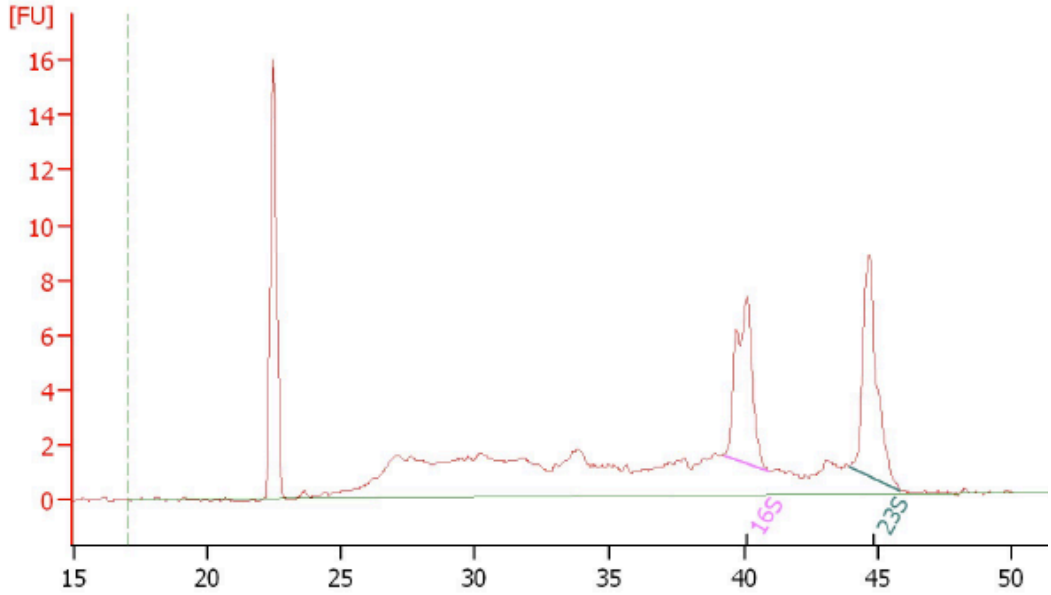


Figure 4.5 Bioanalyzer trace of RNA extracted from string incubated in BM1J leachate (as an example) following small RNA removal and rRNA removal using MICROB*Express* Bacterial mRNA Enrichment kit (Ambion). The 16S and 23S rRNA peaks were found to have been reduced in size without the formation of degradation products.

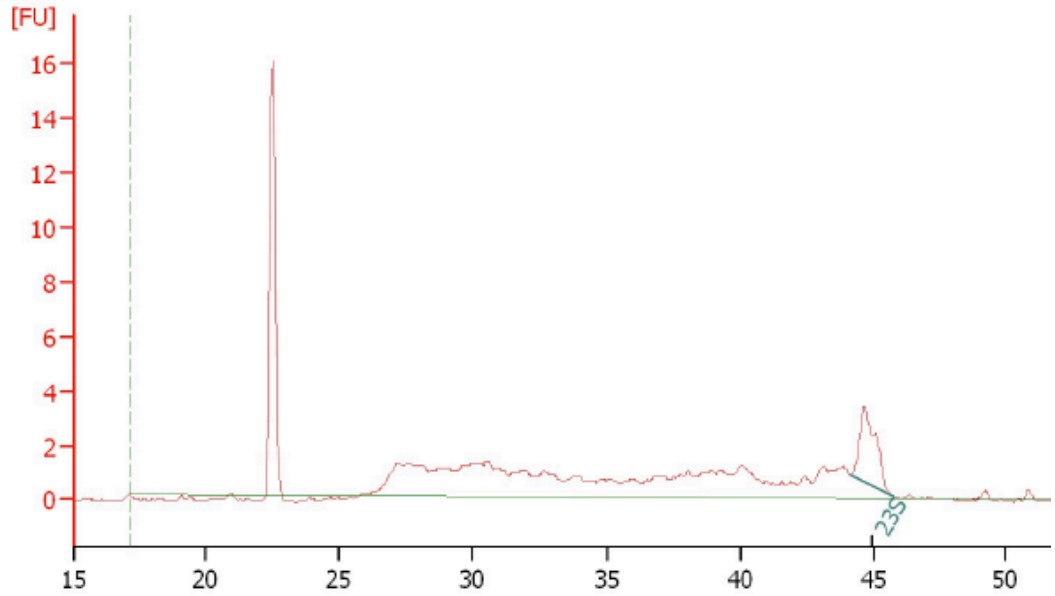


Figure 4.6 Bioanalyzer trace of RNA extracted from string incubated in BM1J leachate (as an example) following small RNA removal and rRNA removal using MICROB*Express* Bacterial mRNA Enrichment kit (Ambion) as well as the Terminator 5'-Phosphate-Depenent Exonuclease (Epicentre). Almost all of the bacterial SSU and LSU rRNA has been depleted.

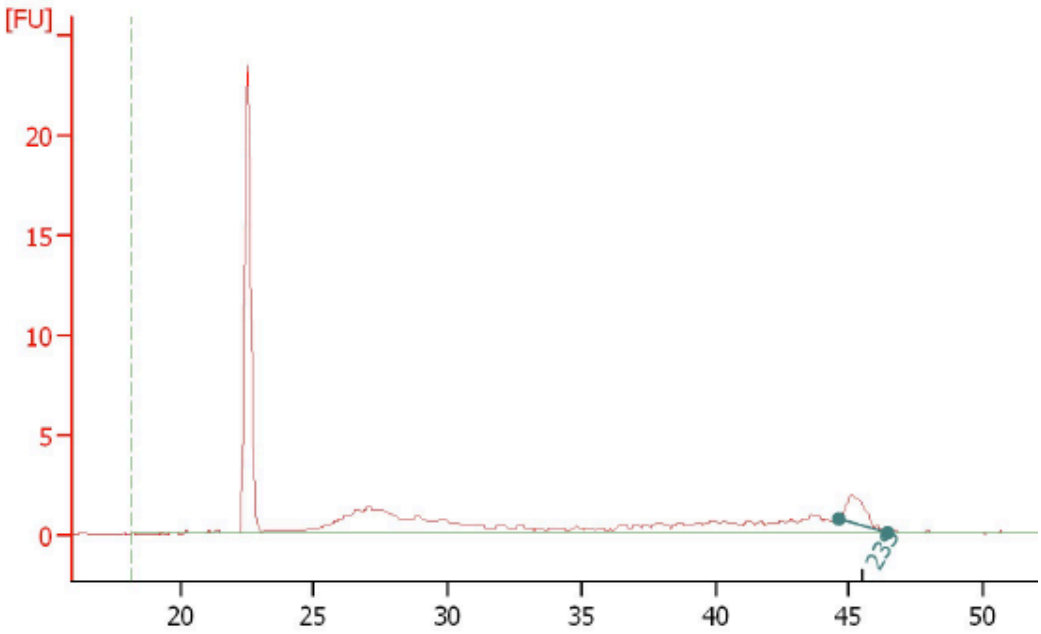


Figure 4.7 Bioanalyzer profile of the enriched mRNA (extracted from colonised string samples BD1, BD2, BD3, BM3E, BM3F, BM3G and BM1J) sent for sequencing on the Illumina MiSeq platform.

4.2.3 Amplification of metatranscriptomic mRNA

MessageAmp II aRNA Amplification kit (Invitrogen) has been widely used in metatranscriptomic studies (Stewart *et al*., 2010; Martinez *et al*., 2010; McCarren *et al*., 2010; Shi *et al*., 2011; Ottesen *et al*., 2011; Stewart *et al*., 2012; Mason *et al*., 2012; Lesniewski *et al*., 2012; de Menezes *et al*., 2012) since the yield of mRNA from precious and troublesome samples is not always adequate for high throughput sequencing. Amplification is achieved as mRNA is reverse transcribed into double stranded cDNA, followed by in vitro transcription for the synthesis of antisense RNA. It was originally planned that the extracted mRNA would be amplified prior to submission for cDNA library preparation. With the technological advancements in Illumina sequencing however, the starting quantity of mRNA required to provide meaningful sequencing output is now very low (< 150 ng per sample). Nonetheless, owing to the popularity of the kit, it was decided that an investigation should be conducted into whether this amplification step introduced any bias into the taxonomic and functional profile of the microbial community under survey.

Generally, a single round of amplification is sufficient to generate microgram amounts of amplified mRNA, the bulk of which is in the size range ~ 1,000 bp. However, it is possible to boost the mRNA yield further by performing a second round amplification if required. It is worth noting that the second round amplification is performed using a different set of primers (included in the kit) and as such, shorter amplification products are produced, typically in the size range 300-500 bp. As the kit is designed for use with mRNA with poly-A tails, an aliquot of the processed, pooled mRNA sample was polyadenylated using the *E. coli* Poly(A) Polymerase enzyme (New England Biolabs) followed by amplification using the MessageAmp II kit (section 2.2.8.6). The T7 oligo(dT) primer provided with the kit was replaced with the T7 BpmI

oligo(dT) primer, as the modified primer contains a BpmI restriction digestion site, which allows for removal of the poly-A tails post-amplification. It was determined that no mRNA was generated from the first round of amplification. Hence, second round amplification was also performed, yielding only ~3.5 μg of amplified mRNA from 250 ng of pooled mRNA from colonised cotton.

Given that Illumina MiSeq can generate a sequencing output of read length up to 500 bp due to its 250x250 bp paired-end chemistry, it was concluded that second round amplification products would not be ideal for such sequencing in order to maximise the potential for drawing meaningful biological conclusions from the data. The kit provides HeLa total RNA to be used as a control, and 1 μg of this was used for amplification with both the T7 oligo(dT) primer and the modified T7 BpmI oligo(dT) primer. Following first round amplification, 72 μg of amplified mRNA was produced using the T7 oligo(dT) primer, whereas only 12 μg of amplified mRNA was produced when using the T7 BpmI oligo(dT) primer. Hence, the use of a modified primer is suggested as one of the reasons for the lack of first round amplification observed when using the kit with the sample mRNA.

In order to improve the yield of first round amplified mRNA from the sample, two modifications were made to the protocol- a hybridisation oven was used for the *in vitro* transcription (Poretsky *et al*., 2009) and the incubation time for the *in vitro* transcription was increased from 14 hours to 18 hours. 70 μg of amplified mRNA was successfully generated form 250 ng of pooled mRNA extracted from colonised cotton. Integrity of the amplified mRNA was assessed using the Bioanalyzer (Fig. 4.8), and it was found to match the profile expected to be generated according to the kit manual. It is suggested that the use of a hybridisation oven is important for optimal *in vitro* transcription as it maintains a uniform temperature inside the whole tube, preventing

the formation of water droplets due to condensation. Despite using a thermal cycler with a heated lid, as recommended by the kit manufacturer, a small amount of condensation was observed to have occurred within the tubes during the *in vitro* transcription step. Preventing even modest amounts of condensation is crucial according to the MessageAmp II aRNA Amplification kit as it can lead to significant changes in the concentration of enzymes and reagents, leading to poor mRNA amplification.

250 ng of mRNA amplified from the mRNA pooled from the cotton was sent to the Centre for Genomic Research, where a ScriptSeq cDNA library was produced before the sample was sequenced on the Illumina MiSeq platform. Paired-end sequencing was performed, with a chosen read length of 250x250 bp.

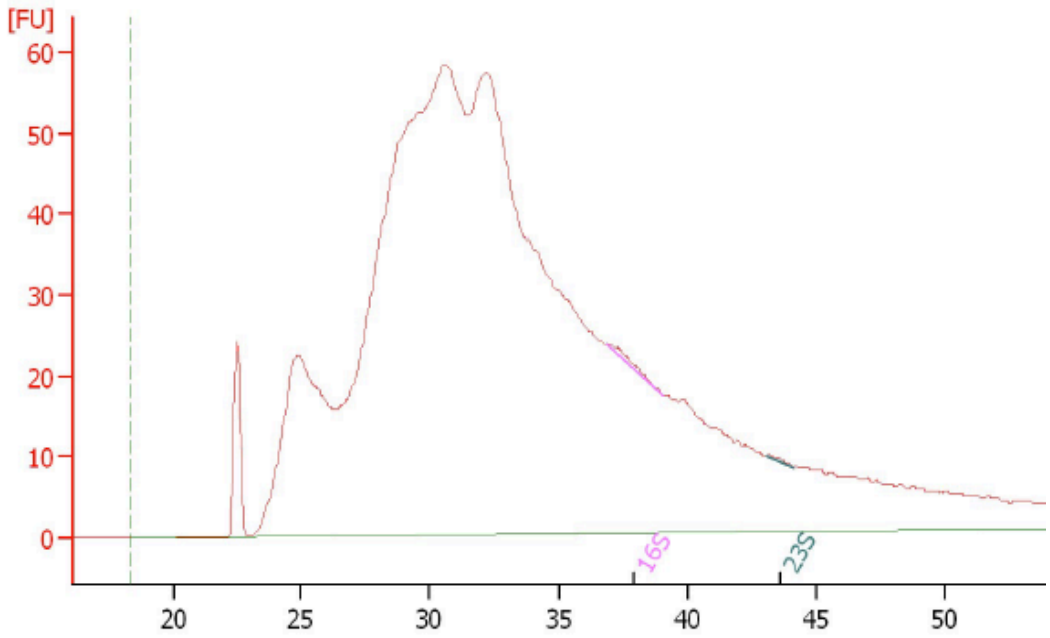


Figure 4.8 Bioanalyzer trace of the amplified mRNA sent for sequencing on the Illumina Miseq platform after first round MessageAmp II aRNA Amplification kit (Invitrogen) amplification. The mRNA profile seen here is typical following a successful amplification using the kit.

4.3 Bioinformatic analyses of the high throughput sequencing data

The sequencing output from the Illumina MiSeq consists of three files: two files consisting of corresponding paired-end reads and a third file with singleton reads. A total of 23.70 million sequences were generated from the metagenome, of which 23.65 million reads were pair-ended. The metatranscriptome and the amplified metatranscriptome datasets consisted of 31.30 million and 33.94 million sequences, of which 30.90 and 33.56 million reads were paired-ended, respectively. Given that the number of singleton reads in all three datasets was very low compared to the number of the paired reads, the former were omitted from analyses in order to ensure that the proportion of the longest possible reads incorporated into the analyses was high.

4.3.1 Quality control and data pre-processing

Mate-paired reads from the three datasets were uploaded on to Galaxy (Afgan *et al*., 2016, accessible at www.usegalaxy.org). Galaxy is an open source, web-based platform that encompasses various tools for next generation sequence data analyses, including quality control and filtering, converting between formats and RNA-seq analysis amongst others. Initial quality check of the data was performed using the tool FastQC to assess the sequencing output prior to any further analysis. FastQC generates graphs and tables for quick data visualisation and evaluation, and a summary of the statistics for the three datasets is provided in Table 4.1. Sequence duplication level was assessed to be high for both the metatranscriptome and the amplified metatranscriptome. However, most of the top duplicated sequences were found to be 16S and 23S rRNA reads, suggesting that rRNA depletion had been far from complete despite the two step removal approach employed. A few reads (<1.0%) were also determined to contain ambiguous bases (called N in read data).

Table 4.1 Summary statistics of the paired-end sequence reads in the metagenome, the metatranscriptome and the amplified metatranscriptome datasets generated using FastQC tool on the Galaxy webserver.

	Metagenome	**Metatranscriptome**	**Amplified Metatranscriptome**
Mean read length (bp)	238	168	159
Mean Phred base quality score	36.6	36.0	35.6
GC content (%)	52	51	51
Per base N content (%)	<1.0	<1.0	<1.0
Sequence duplication level (%)	5.7	75.6	78.2

Raw data were subjected to stringent quality control in order to yield high quality sequences for downstream analyses using the Filter FASTQ tool available on Galaxy (Blankenberg *et al*., 2010). Based on the sequence quality information produced by FastQC, the following criteria were applied for extensive filtering of the sequence reads:

- Remove sequences with length <125 bp (<150 bp for the metagenome)
- Remove sequences with mean Phred base quality score <25
- Remove artificial duplicate sequences
- Remove sequences with more than 1 ambiguous base (N)

A Phred score of 25 equates to less than 1% chance of the base being miscalled and was selected as the threshold for filtering of sequences (Kunin *et al*., 2008). Filter FASTQ only removed true artificial duplicates and not the SSU and LSU rRNA sequences that made up the vast majority of the sequences flagged as artificial replicates. Furthermore, while 85% of all sequences passed the filtering criteria from the metagenomic dataset, a relatively lower 66% and 59% of the reads were retained from the metatranscriptome and the amplified metatranscriptome, respectively. This suggests that while the mean Phred base quality score was high for these two datasets (Table 4.1), the number of sequences with mean quality score just under the threshold value of 25 was significant, leading to their removal. Despite submission of high quality processed mRNA for sequencing, the average read length was lower and the proportion of reads with low quality score bases was higher in the two metatranscriptomic datasets compared to the metagenome. The Centre for Genomic Research suggested that similar results had been obtained when metatranscriptomes derived from heterogeneous microbial communities were sequenced from environmental samples in the past.

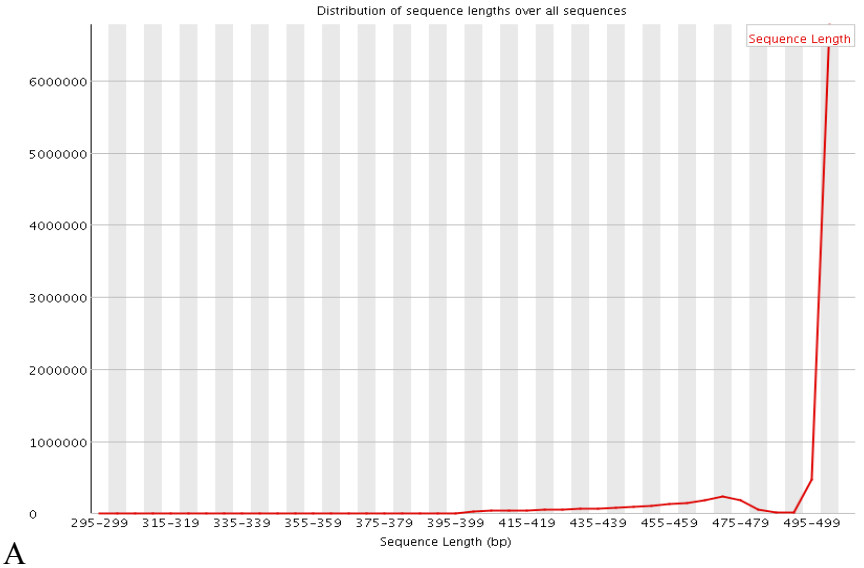
Assembly of the quality filtered sequences was not attempted as coverage of a significant proportion of the metagenome and the metatranscriptomes was expected to be insufficient, while a considerable amount of genetic information would still be left in unassembled reads even if contigs could be produced. Moreover, majority of the contigs generated would be short due to a number of sequences likely to be represented only once or twice, rendering comprehensive assembly impossible without a reference metagenome or metatranscriptome to map against. Short sequence aligners such as Bowtie (Langmead *et al*., 2009) and Bowtie2 (Langmead & Salzberg, 2012) usually require greater sequencing depth than afforded by the Illumina MiSeq. Detailed analysis of gene expression allowing for a comparison to be drawn between the metagenome and the metatranscriptome would also not be possible following assembly.

Instead, the paired reads that passed stringent quality control were combined using the FASTQ Joiner tool on Galaxy (Blankenberg *et al*., 2010), yielding longer reads that allow for more robust analyses to be performed downstream. A total of 9.07 million, 9.59 million and 8.58 million full length sequences were generated from the metagenome, the metatranscriptome and the amplified metatranscriptome, respectively. FastQC analysis was performed on the paired sequences for all three datasets to generate summary statistics, and read length distribution is presented in Figure 4.9. 80% of the total sequences in the metagenomic dataset were determined to be 495-500 bp, and that number dropped to 24% and 16% for the metatranscriptome and the amplified metatranscriptome, respectively. While that number seems low for the metatranscriptomes, it is worth noting that only 12-15 months before the samples were sequenced, paired-end sequencing on the Illumina MiSeq generated reads of only 150x150 bp meaning that a maximum length of 300 bp was achievable. Mean read length of 489 bp, 386 bp and 373 bp in the metagenome, the metatranscriptome and the

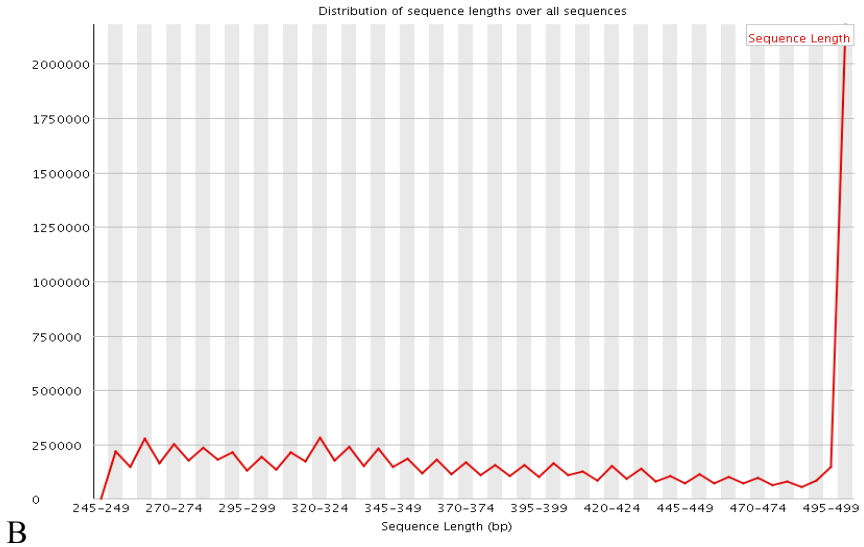
amplified metatranscriptome, respectively, shows a marked improvement in sequencing output read length.

A total of 99.8% of sequences in the metatranscriptome and 99.4% of the sequences in the amplified metatranscriptome had a mean Phred quality base score >=33, while only 202 out of 9.07 million sequences in the metagenome did not meet that criteria. A larger chunk of sequences was discarded from the two metatranscriptomic datasets following stringent quality filtering, which ensures that only high quality data were used for downstream analyses.

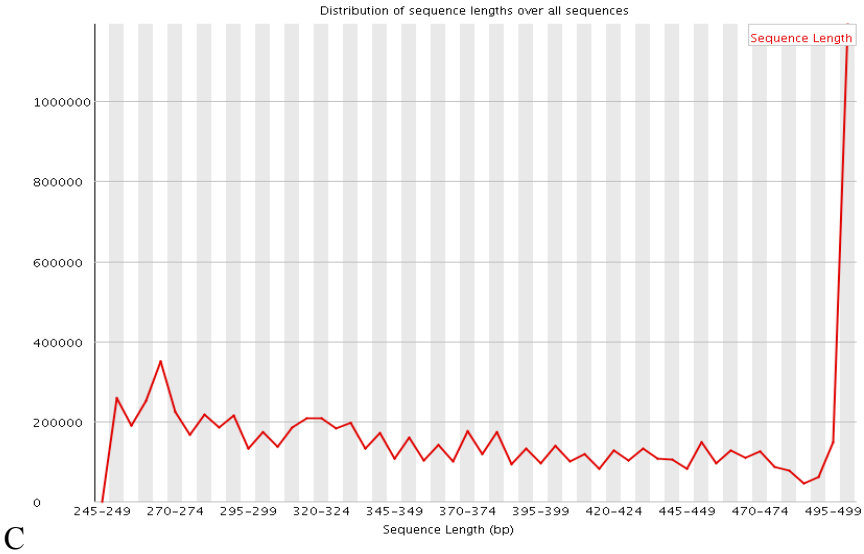
The free to use service and an easy to use interface makes Galaxy an attractive alternative to coding using Perl and Python for execution of bioinformatic applications using command line. Its modular nature also makes it a very useful tool as research groups can create their own custom analysis workflows by adding further analysis tools from the Galaxy tool shed. While most custom Galaxy platforms are private, some servers are publically available (accessible at www.galaxy.nbic.nl and www.mississippi.snv.jussieu.fr amongst others), and incorporate tools such as those required to perform multiple alignment, mapping as well as Blast against custom databases (for sequences in the range of thousands). However, computationally intensive tasks involving large datasets, such as Blast on millions of sequence reads against NCBI nr database, are usually unable to be performed on Galaxy due to limitations with processing power.



A



B



C

Figure 4.9 Graphs demonstrating read length distribution of joined paired-end reads in the metagenome (A), the metatranscriptome (B) and the amplified metatranscriptome (C) datasets generated using FastQC tool on the Galaxy webserver.

4.3.2 Taxonomic and functional analyses using MG-RAST

High quality sequences from the three datasets were uploaded onto MG-RAST webserver. MG-RAST has a growing collection of over 260,000 metagenomes, which it incorporates into its M5NR database. It processes data through its own analysis pipeline in order to predict protein-coding ORFs and rRNA features, and is particularly useful for providing an overview of microbial community composition as well as a breakdown of functional genes present in the dataset. Such an approach allows for comparative analysis of the datasets as the metagenome provides a general summary of the overall gene content of the community, while the metatranscriptome offers an insight into genes being actively expressed in the same environment.

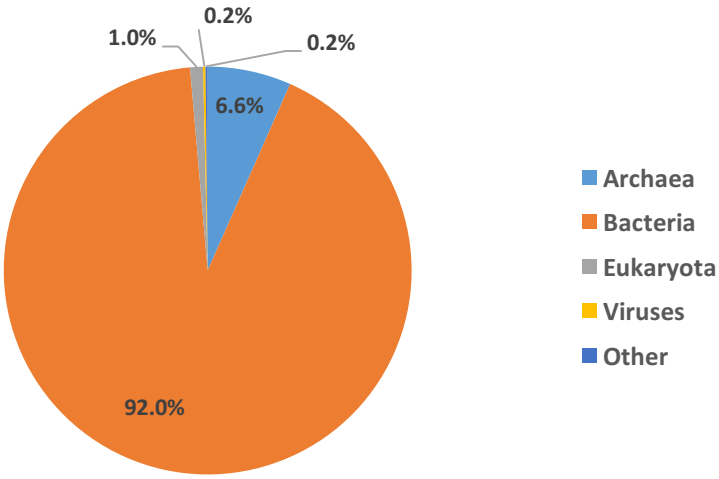
Every dataset uploaded onto MG-RAST is subjected to its own built-in quality control measures, including filtering of sequences based on length, Phred quality score and artificial duplicate removal. Sequences with a high similarity to those in SILVA (Pruesse *et al*., 2007), Ribosomal Database Project (RDP) (Cole *et al*., 2007) and Greengenes (DeSantis *et al*., 2006) rRNA databases could also be removed. Since the uploaded sequences had already been quality filtered based on length and quality scores, the vast majority of sequences removed were those identified as rRNA sequences and artificial replicates. 5.1% of metagenomic reads, 80.1% of the metatranscriptomic reads and 85.4% of the amplified metatranscriptomic reads were omitted from the analyses. The remaining sequences were subsequently queried against the M5NR database.

4.3.2.1 Domain-level taxonomic classification

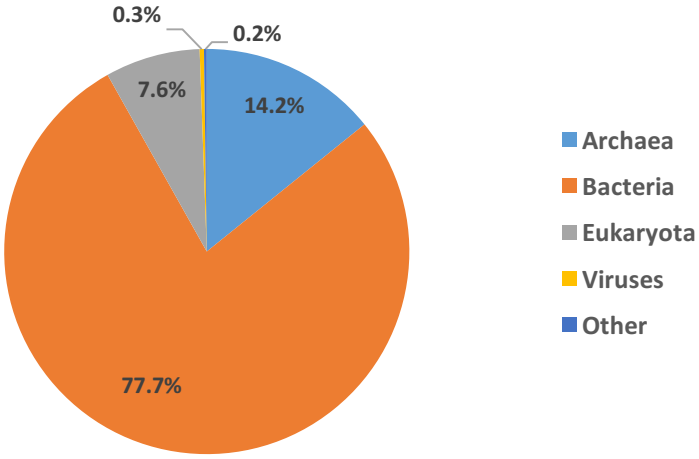
A summary of the microbial community composition at the domain level in the metagenome, the metatranscriptome and the amplified metatranscriptome is presented

in Figure 4.10. The metagenome is dominated by Bacteria as they accounted for 92% of the total reads, followed by Archaea (6.6%), Eukaryota (1%) and Viruses (0.2%) (Fig. 4.10A). Reads that could not be assigned to a domain due to their unknown identity were classed under the category 'Other', and these contributed 0.2% of the biological diversity in the metagenome. While Bacteria, Archaea and Eukaryota were determined to be the three most populous domains in the metatranscriptome as well, there was a marked shift in their overall abundance (Fig. 4.10B) when compared to the metagenomic data. While the percentage of reads assigned to Bacteria decreased to 77.7%, those assigned to Archaea and Eukaryota increased to 14.2% and 7.6%, respectively. This suggests that, although smaller in size than Bacteria, the Archaea and Eukaryota communities are highly active in the environment under survey. Reads assigned to Viruses and Other remained almost identical.

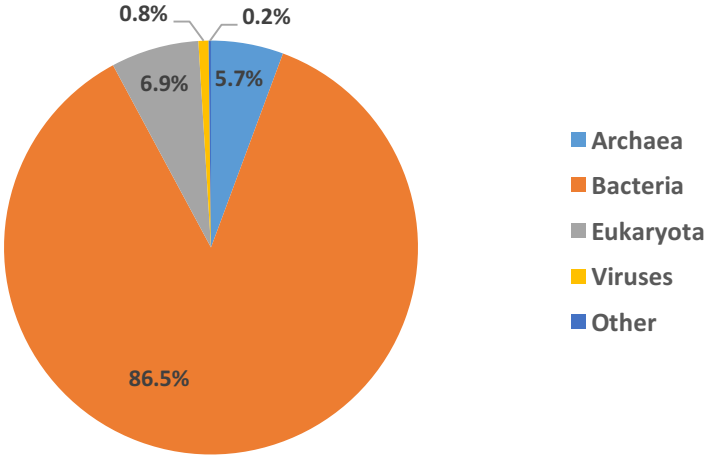
Comparison of the amplified metatranscriptome community profile with that of the metatranscriptome suggests that Bacteria are dominant again as their abundance rises to 86.5%. It is notable, however, that while the abundance of Eukaryota and Viruses remains roughly the same, the percentage of reads assigned to domain Archaea drops significantly to 5.7% (Fig. 4.10C). It is worth mentioning that any unclassified sequences that were deemed to be derived from Archaea, Bacteria, Eukaryota, Fungi or Viruses were reported back as assigned to their respective domains at this level of classification for all datasets.



A



B



C

Figure 4.10 Pie charts demonstrating the microbial community composition at the domain level within the (A) metagenome, (B) metatranscriptome, and (C) amplified metatranscriptome generated following MG-RAST analysis.

4.3.2.2 Phylum-level classification

Microbial community composition was analysed at the phylum level for a more considered analysis of the most abundant microbes in the high throughput sequencing datasets. It is worth noting that annotation of metatranscriptomic reads points towards the transcriptional activity of microbes rather than a presence/absence analysis, and hence provides an insight into which organisms are the most proactive in a community. Abundance (expressed as percentage of total reads assigned) for the most populous phyla as well as phyla containing cellulolytic microbes of interest is presented in Figure 4.11, allowing for a comparative examination of the three datasets. Reads listed as Unclassified from the different domains are also displayed. Proteobacteria, Firmicutes, Bacteroidetes, Chloroflexi, Planctomycetes and Actinobacteria contain the most bacterial representatives in the metagenome, with Proteobacteria (30.8%) and Firmicutes (27.6%) accounting for almost 60% of the phylogenetic diversity. Phylum Euryarchaeota (6.2%) had the most archaeal reads assigned. Presence of a high number of Firmicutes in leachate is unsurprising, as the phylum contains class Clostridia, and representatives would be expected to be numerous in an anoxic environment rich in organic carbon. Moreover, several members of this obligate anaerobic class of microbes have been reported to have an active role in lignocellulose decomposition in landfill. Phylum Bacteroidetes also contains cellulolytic bacteria such as Cytophaga spp. (Lynd *et al*., 2002). Fibrobacteres made up 0.1% of the microbial community with 7,119 reads assigned to the phylum.

Proteobacteria (26.4%) and Firmicutes (23.9%) also dominated the phylum level biodiversity in the metatranscriptome. However, a few interesting outcomes were evident when the metatranscriptome was compared with the metagenome. There was a small drop in the abundance of Proteobacteria, Firmicutes and Bacteroidetes, perhaps

suggesting that cellulose decomposition in landfill is not as heavily influenced by the action of cellulolytic genera in these phyla as might have been suggested by metagenomic analysis alone. Phyla Chloroflexi and Cyanobacteria consist of phototrophic forms (Garrity & Holt, 2001), and a decrease in their abundance in the metatranscriptome is not surprising. Unclassified sequences comprise ~ 11% of total microbial sequences in the metatranscriptome compared to only 1.2% in the metagenome, suggesting that a significant number of never sequenced before microbes are active in the environment and might have a role in cellulose colonisation and hydrolysis. A marked increase in the percentage of reads assigned to Euryarchaeota is expected, as the phylum consists of methanogenic Archaea such as members of the genera *Methanococcus*, *Methanobacteria*, *Methanolobus* and *Methanosarcina* (Lynd *et al*., 2002), and leachate is an actively methanogenic environment.

Comparison between amplified and non-amplified metatranscriptomes suggested a decrease in the abundance of Euryarchaeota, Planctomycetes as well as Unclassified (derived from Archaea and Bacteria) in the latter. Furthermore, a small increase in Firmicutes and curiously, a large increase in the abundance of Proteobacteria was also observed. Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria comprise of a large number of representative species that inhabit a variety of environments, including soil (Tveit *et al*., 2013) and the rumen (Brulc *et al*., 2009). As such, these phyla have been reported to account for a major proportion of metagenomic and metatranscriptomic datasets (Xiong *et al*., 2012). The majority of studies addressing microbial community diversity in landfill leachate have been based on SSU rRNA gene amplicon sequencing, and predominance of phyla Proteobacteria, Bacteroidetes and Firmicutes within domain Bacteria and Euryarchaeota within domain Archaea has been widely reported (Stamps *et al*., 2016; Yadav *et al*., 2015). This study

is the first incidence of microbial community profiling in landfill leachate combining both WGS metagenomics as well as metatranscriptomics. A study involving WGS metagenomic analysis of the microbial community colonising cotton 'bait' incubated inside landfill leachate also reported Firmicutes, Bacteroidetes and Proteobacteria to be the most abundant bacterial phyla (personal communication with Dr E. Ransom-Jones, Bangor University).

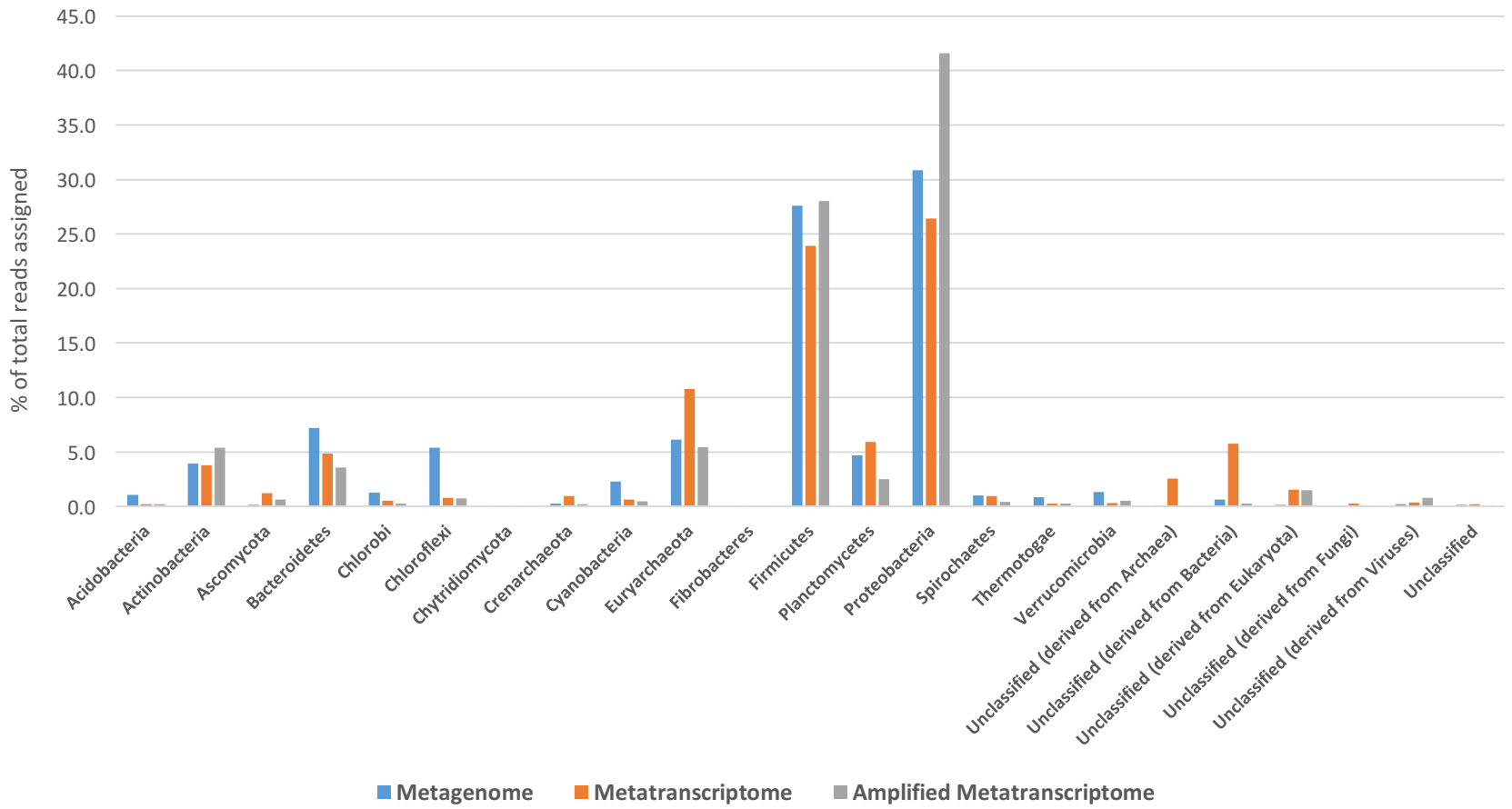


Figure 4.11 Bar chart showing the microbial abundance (expressed as percentage of total reads assigned) for the most populous phyla as well as phyla containing cellulolytic microbes of interest in the metagenome, the metatranscriptome and the amplified metatranscriptome generated following MG-RAST analysis.

4.3.2.3 Phylum-level eukaryotic classification

Given our interest in the eukaryotic component of the microbial community involved in cellulose hydrolysis following on from the unsuccessful attempt to culture anaerobic cellulolytic fungi and the subsequent molecular cloning and 454 pyrosequencing data analysis (chapter 3), taxonomic classification of reads assigned to the domain Eukarya in the metagenomic and metatranscriptomic data was assessed in more detail. Abundance of the most populous eukaryotic phyla as well as phyla containing cellulolytic microbes of interest is presented in Figure 4.12 as a percentage of total reads assigned to domain Eukaryota.

Streptophyta, Chordata, Ascomycota, Unclassified (derived from Eukaryota) and Arthropoda sequences dominate the metagenome. Perhaps disappointingly, Neocallimastigales accounted for only 0.06% (63 reads) and 0.003% (5 reads) of the eukaryotes in metagenome and metatranscriptome, respectively. However, Chytridiomycota comprised 0.3% of eukaryotic reads in the metatranscriptome, with a rise from 1 assigned sequence in the metagenome to 546 in the metatranscriptome, suggesting that these anaerobic fungi are active in the sample. Interestingly, the percentage of assigned unclassified reads accounted for ~ 14% of the metagenome and ~ 23% of the metatranscriptome, as an increase in reads derived from both Eukaryota and Fungi was observed.

Closer inspection of the unclassified reads derived from Eukaryota revealed that this category was composed almost entirely of ciliated and flagellated protozoal reads. An increase in abundance suggests that these microbes could either be involved in cellulose hydrolysis in landfill or in active grazing of bacteria in the leachate. Abundance of Family Litostomatea, members of which have been well documented to be important lignocellulosic biomass degraders in the herbivore rumen (Leschine,

1995; Lynd *et al*., 2002), was determined to be 0.01% and 0.02% of all Eukaryota reads in the metagenome and the metatranscriptome, respectively. More fascinatingly however, Family Metopidae accounted for 0.003% of Eukaryota and 0.2% of the Unclassified (derived from Eukaryota) reads in the metagenome. The abundance of this Family, which includes genera with a reported role in cellulose degradation such as *Metopus* and *Nyctotherus* (Gijzen *et al*., 1994; Lynn & Wright, 2013) increased to 1% of all Eukaryota and 7% of all Unclassified (derived from Eukaryota) reads in the metatranscriptomic data, suggesting that these microbes are active in the environment under survey. Their grouping under unclassified microbes reiterates the fact that there is practically no data available on the role, or indeed occurrence, of ciliated protozoa in landfill.

Despite performing enrichment using the dewaxed cotton bait, a significant proportion of the taxonomic diversity consists of microbes not directly involved in cellulose hydrolysis. This comes as no surprise as the cotton has been incubated in the leachate and enough of the liquid covering the baits was carried over along with the microbial biofilm on the cotton while performing nucleic acid extractions. Leachate in the two landfill sites is comprised of household waste, sludge, paper mill waste etc. and that explains the presence of a large number of reads from phylum Chordata in the metagenome. However, the number drops significantly in the metatranscriptome, as those cells are not expected to be active in such an environment. Moreover, the leachate collected from Bidston Moss site was found to contain a large amount of plant and animal matter as the sampling boreholes were left uncovered in some instances. This potentially explains the large number of Chlorophyta and Streptophyta sequences found in the metagenome, as these phyla harbour higher plants and algae (Bremer, 1985). Their abundance does indeed decrease in the metatranscriptome. Elsewhere, there was

an increase in the percentage of reads assigned to Nematoda and Arthropoda, and a slight decrease in the percentage of reads assigned to Basidiomycota in the metatranscriptome.

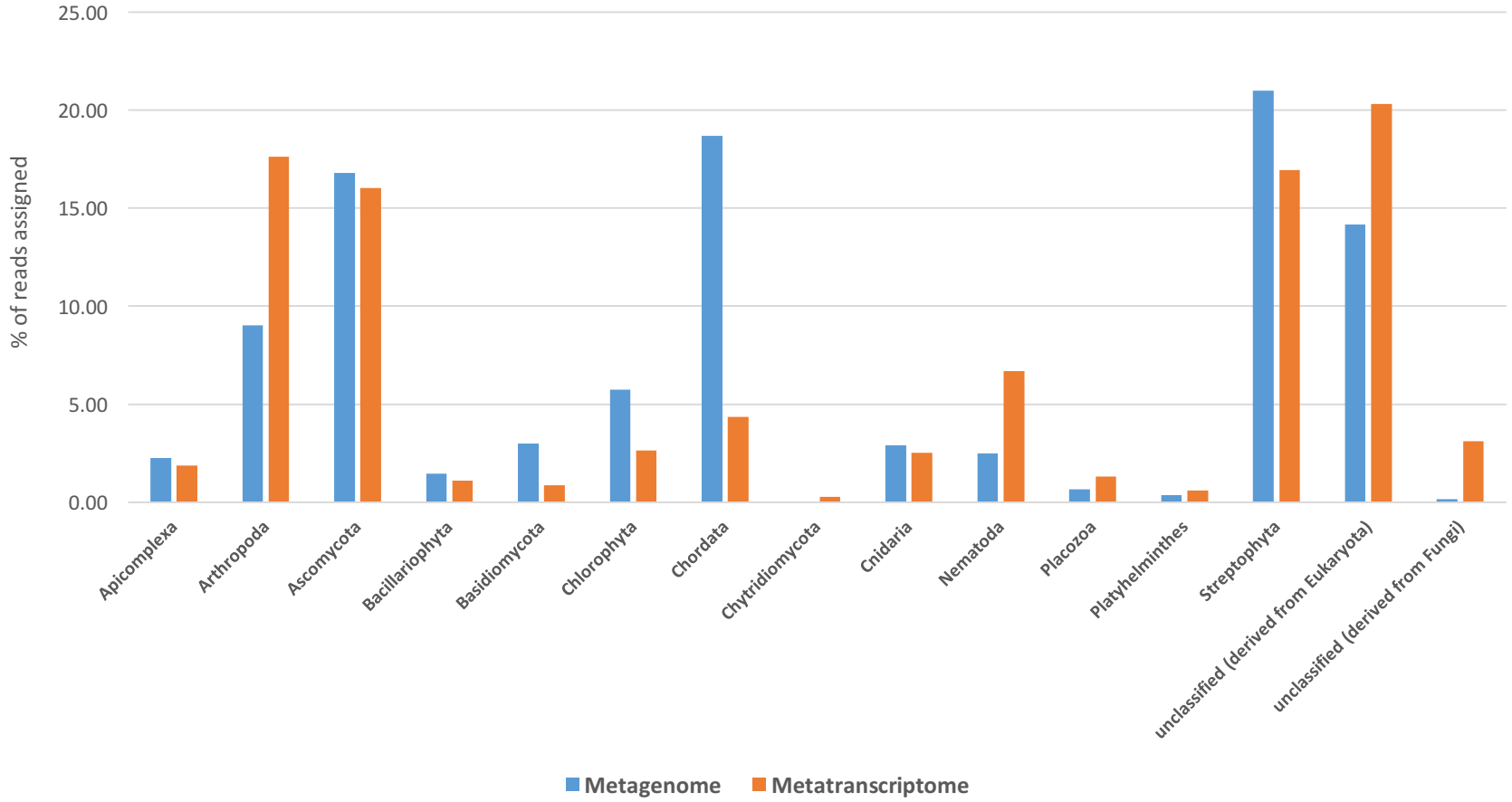


Figure 4.12 Bar chart showing the eukaryotic abundance (expressed as percentage of total reads assigned to domain Eukaryota) for the most populous phyla as well as phyla containing cellulolytic microbes of interest in the metagenome and the metatranscriptome generated following MG-RAST analysis.

4.3.2.4 Functional annotation using SEED subsystems

A summary of annotated proteins from the metagenomic and metatranscriptomic datasets that could be assigned to a SEED subsystems (Overbeek *et al*., 2005) category is provided in Figure 4.13, facilitating a comparative analysis into the expression profile of the biofilm microbial community. The metagenome is dominated by reads with an apparent function in Clustering-based subsystems, Carbohydrate metabolism, Protein metabolism and metabolism of Amino acids and derivatives, while most reads in the metatranscriptome could be assigned to functional categories involving RNA metabolism, Clustering-based subsystems, Protein metabolism and Carbohydrate metabolism. This, along with a significant number of reads also involved in Membrane transport, Respiration, Lipid metabolism as well as Nucleoside and nucleotide based subsystems points towards an active microbial community where mRNA transcripts and protein processing represents a dynamic cellular lifestyle.

It also appears that transposable elements, including phage, prophages and plasmids are active, which is to be expected in a mixed microbial community extracted from an environment under selection pressures that include anoxia amongst others. Comparison of the two metatranscriptomes suggested that reads annotated with a function in protein metabolism and stress response were slightly higher, while those involved in DNA and RNA metabolism were slightly lower in the amplified metatranscriptomic dataset.

Interestingly, 0.9% and 0.4% of total hits with an annotated function in carbohydrate metabolism in the metagenomic data corresponded to glycoside hydrolases and structural components of cellulosome, respectively, and this figure was determined to be 0.6% and 0.3%, respectively, in the metatranscriptomic data.

4.3.2.5 Functional annotation using COG and KEGG orthologies

A brief overview of annotated protein-coding transcripts in the three datasets that could be assigned to a functional category through Clusters of Orthologous Groups (COG, Tatusov *et al*., 2003) and Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa & Goto, 2000) orthologies is provided in Figures 4.14 and 4.15, respectively. COG analysis suggests that the highest percentage of reads in all three datasets were involved in Metabolism. While a small increase in Information Storage and Processing and a small decrease in Metabolism were observed in the metatranscriptome compared to the metagenome, no other significant changes were noticeable when the datasets were compared. 6.9%, 5.8% and 4.8% of the total assigned hits in the metagenome, the metatranscriptome and the amplified metatranscriptome, respectively, were investigated to be directly involved in carbohydrate metabolism. It is also worth mentioning that a significant proportion of reads (17.1% in metagenome, 16.0% in metatranscriptome and 14.3% in amplified metatranscriptome) were determined to be poorly characterised, which is unsurprising given the environmental origin of the sample.

KEGG analysis produced results very similar to those obtained from COG analysis. Unsurprisingly, reads involved in Metabolism formed the dominant category followed by those involved in Genetic Information Processing, while reads involved with a potential role in Human Diseases were determined to be very low. An increase in Information Storage and Processing and a decrease in reads corresponding to Metabolism were observed when the metatranscriptome was compared to the metagenome. A similar effect was observed when the amplified metatranscriptome was compared to the non-amplified metatranscriptome. 13.6%, 11.9% and 9.0% of the total

assigned hits were involved in carbohydrate metabolism in the metagenome, the metatranscriptome and the amplified metatranscriptome, respectively. Interestingly, 2% of the hits assigned to carbohydrate metabolism were annotated as encoding endoglucanases in the metagenome. This number was observed to be 2% and 0.4% in the metatranscriptome and the amplified metatranscriptome, respectively.

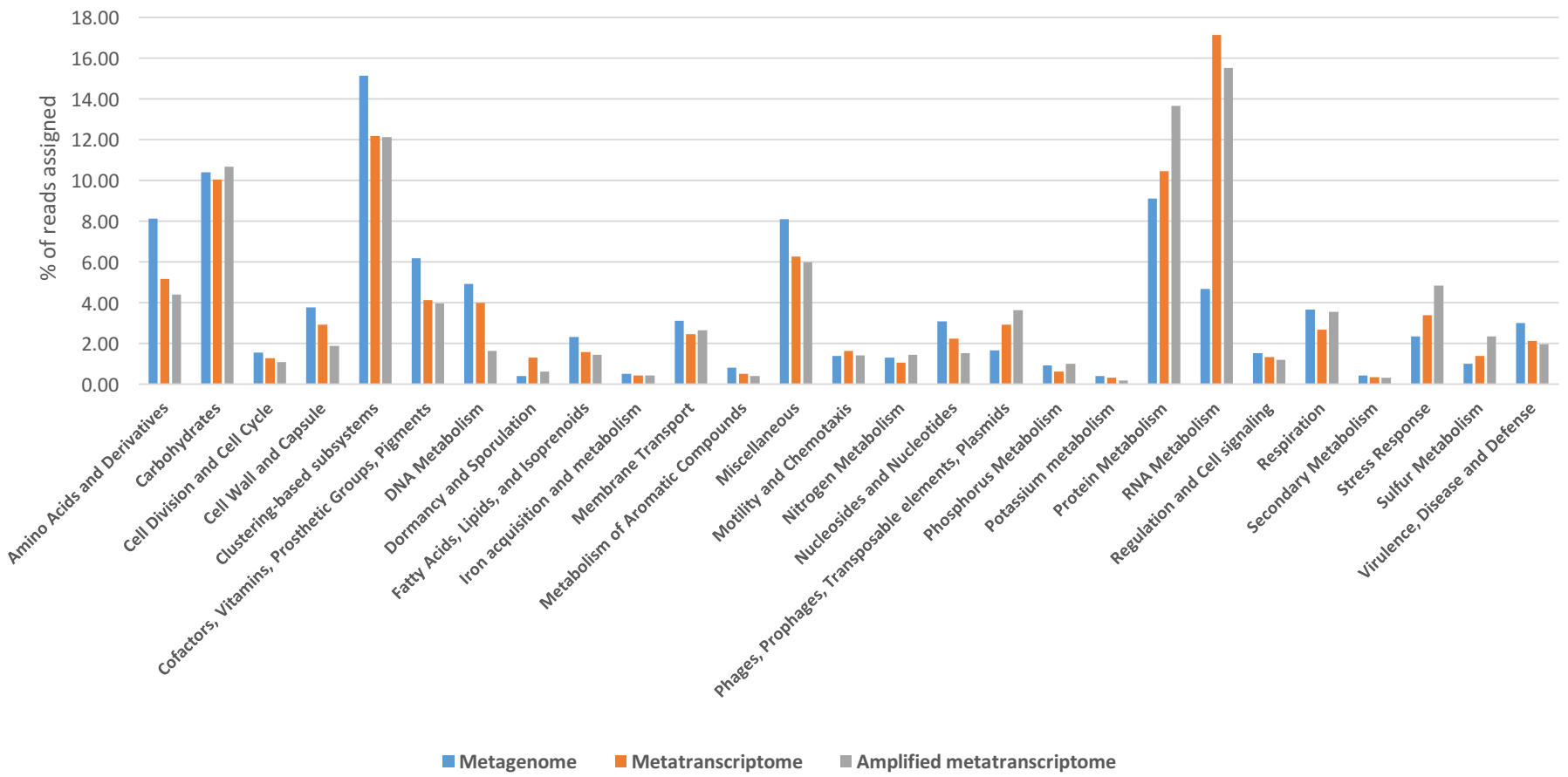


Figure 4.13 Bar chart showing the distribution of annotated proteins as a percentage of reads assigned to SEED subsystems categories in the metagenome, the metatranscriptome and the amplified metatranscriptome, following functional analysis performed using MG-RAST.

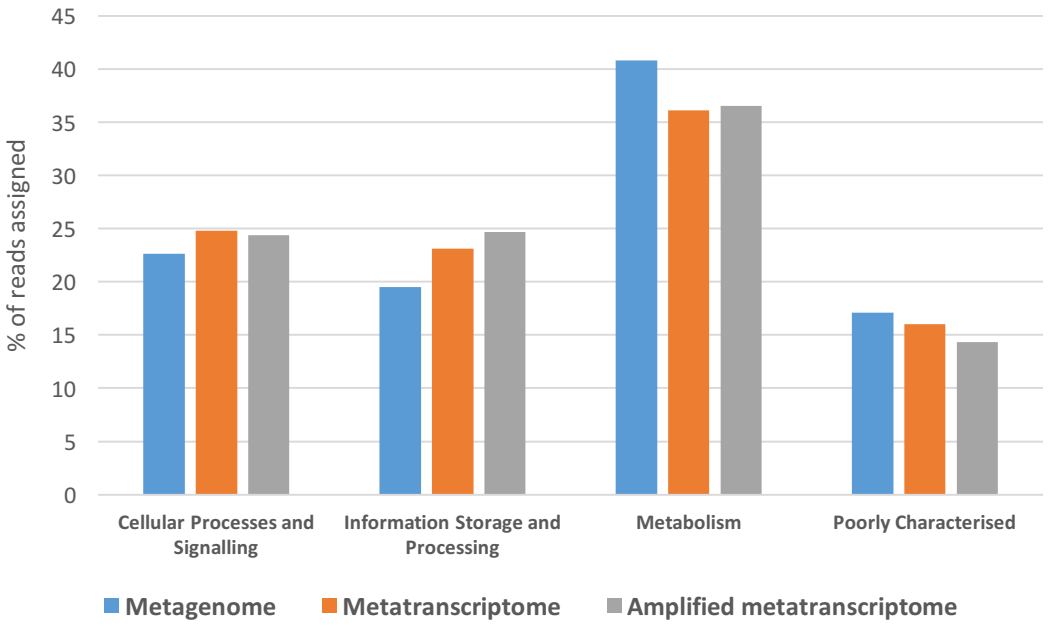


Figure 4.14 Bar chart showing the distribution of annotated proteins as a percentage of reads assigned to COG categories in the metagenome, the metatranscriptome and the amplified metatranscriptome, following functional analysis performed using MG-RAST.

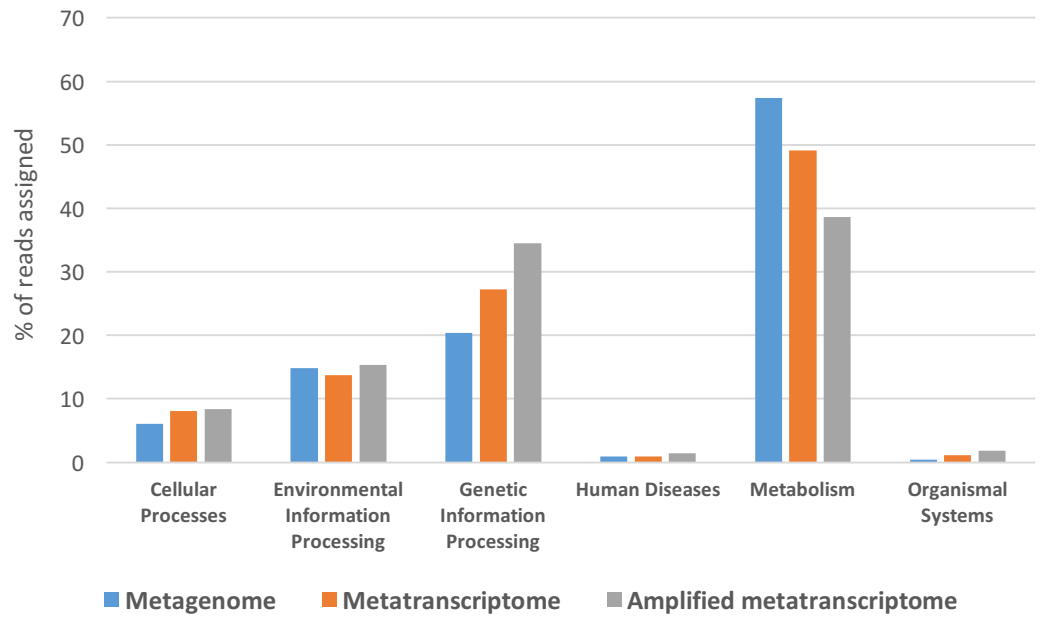


Figure 4.15 Bar chart showing the distribution of annotated proteins as a percentage of reads assigned to KEGG categories in the metagenome, the metatranscriptome and the amplified metatranscriptome, following functional analysis performed using MG-RAST.

4.3.3 Further taxonomic analyses

It is now generally acknowledged that analysis performed using different databases can lead to slightly different annotation of the same dataset (Aguiar-Pulido *et al*., 2016; Bashiardes *et al*., 2016). With this is mind, it was decided that analysis of the community structure and composition from the metagenomic and metatranscriptomic data should be performed using analytical tools that search against separate databases using their own unique algorithms.

4.3.3.1 rRNA removal using SortMeRNA

Before further taxonomic classification could be performed, it was critical to remove reads corresponding to rRNA sequences from the amplified and non-amplified metatranscriptomes. MICROB*Express* Bacterial mRNA Enrichment kit (Ambion) was utilised for rRNA depletion during sample processing, and it uses probe hybridisation to remove bacterial rRNA sequences. However, the probes are designed based on known rRNA sequences in the database, and as such can lead to impartial rRNA removal from prokaryotes that have yet to be classified leading to a skewed microbial diversity analysis. Hence, it was essential to decipher microbial taxonomy based solely on mRNA transcripts. No such issues were expected in the metagenome, however, and rRNA removal was not sought for this dataset.

SortMeRNA (Kopylova *et al*., 2012) was used to perform rRNA sequence removal from the two metatranscriptomic datasets. Quality filtered paired reads were queried against the SILVA SSU (16S and 18S rRNA) as well as LSU (23S and 28S rRNA) database (Pruesse *et al*., 2007) and those with a strong similarity to ribosomal RNA sequences were removed. While 71.6% of all metatranscriptomic reads were removed, 81.25% of all reads in the amplified metatranscriptome were scrapped. The

percentage of sequences mentioned here are roughly similar to those classed by MG-RAST quality control as rRNA sequences or artificial replicate sequences (80.1% and 85.4% in the metatranscriptome and the amplified metatranscriptome, respectively). In both the cases, the amplified metatranscriptome was reported to consist of a larger number of rRNA and artificial replicate sequences.

4.3.3.2 Taxonomic analysis using Kraken and MetaPhlAn

Following rRNA sequence removal, 1.61 million amplified metatranscriptomic and 2.72 million non-amplified metatranscriptomic reads, along with the 9.07 million quality filtered metagenomic reads, were used for taxonomic analysis. Metagenomic Phylogenetic Analysis (MetaPhlAn) compares high throughput sequences against its database of marker genes established from reference genomes, allowing high speed and unambiguous clade-specific taxonomic annotations (Segata *et al*., 2012). It uses BowTie2 to perform initial alignment of the sequences, and is designed to deal with very short sequencing reads (as short as 40 bp). This tool was found on the Galaxy webserver www.orione.crs4.it, and the three large datasets were used to assess its performance. Curiously, the annotation was determined to be as 100% Unclassified for all three datasets. It would appear that while this tool has been used successfully for analysis of metagenomic reads from the human microbiome (Huttenhower *et al*., 2012), sequences from a complex environmental sample such as landfill had no reference sequence in the MetaPhlAn database of marker genes.

Kraken (Wood & Salzberg, 2014) is another taxonomic sequence classification system that is designed to work with short metagenomic sequences ($\geq$100 bp) for very high speed annotation. Kraken scans the k-mers within metagenomic reads, followed by querying of said k-mers against its database which consists of reference k-mers

derived from its own genomic library, allowing for generation of taxonomic labels for the reads. Kraken offers two databases: default database (75 GB size) and MiniKraken (4 GB size). In order to perform taxonomic analysis, the chosen database needs to be pre-loaded into the computer's RAM for quick operation. Given the lack of availability of the crippling computational power required to run Kraken with the default database, MiniKraken database was downloaded along with the Kraken scripts and analysis of the three datasets was performed using command line operation as detailed in the Kraken manual.

The overwhelming majority of reads were annotated as Unclassified, with only 3.21%, 8.36% and 15.85% of the sequences in the metagenome, the metatranscriptome and the amplified metatranscriptome able to be assigned a classification. Since MiniKraken database is considerably smaller than the default database, it only contains a small subset of the most common k-mers present in the larger Kraken database of reference k-mers. Given that MiniKraken operation is extremely rapid and produced classification slightly better than MetaPhlAn, it is likely that using the default database could generate a reliable taxonomic annotation from a heterogeneous microbial community. Such an operation would only be possible through the use of a large computing cluster with processing power of at least 80 GB RAM. The inability to generate any meaningful community composition overview using the reference sequence-based MetaPhlAn and Kraken classification systems perhaps suggests that a significant proportion of the microbial community under survey here is yet to be classified.

4.3.4 Mining for cellulases in metatranscriptomic data

One of the primary objectives of the project was to profile the cellulases being expressed by the mixed microbial community in the landfill leachate. In order to do so, the 2.72 million quality filtered and rRNA removed sequences in the metatranscriptomic dataset were uploaded on to the MetaGeneMark webserver (Zhu *et al*., 2010) for open reading frame (ORF) prediction. MetaGeneMark has been optimised for gene prediction in metagenomic reads, and such an approach has been used successfully by Hess *et al*. (2011), as well as by Dr. James Houghton in our research group for identification of reads corresponding to glycoside hydrolases (GHs) from high throughput metagenomic and metatranscriptomic data (Houghton 2013, p. 154). A total of 2.34 million genes were predicted by MetaGeneMark in the metatranscriptome.

4.3.4.1 Assigning sequences to glycoside hydrolase families using Pfam database

The predicted ORFs were used as an input to query against the high-quality, curated protein family database Pfam (Finn *et al*., 2010). The large Pfam-A database was downloaded, and the search was performed locally using the pfamscan.pl script after installation of the HMMer3 software as following instructions from Pfam. The method allows for identification of ORFs with a high similarity to GH families as determined by the Pfam database, and as such reveals mRNA transcripts that partially encode enzymes involved in hydrolysis of polysaccharides, including cellulose and hemicellulose. A total of 1,724 ORFs were determined to have a high similarity to a GH family in the Pfam database.

The distribution of ORFs with hits to GH families as determined by Pfam is presented in Fig. 4.16. GH families 9, 2, 5, 10 and 3 were the only ones with greater

than 100 reads assigned to them, with the highest number of hits assigned to GH family 9 (151). The hits presented as GH family 5 were referred to as Cellulase by the Pfam database, which is a misnomer, as GH family 5 was initially known as cellulase family A according to the Carbohydrate-Active enzymes database (CAzy). This is due to the fact that earlier classification and grouping of enzymes in GH families was based on their substrate specificity. This has since been altered, and classification into GH families is now based on amino acid sequence similarity. As such, enzymes classed within each GH family can display specificity to multiple substrates, and classification of a read into a GH family in not a conclusive indication of catalytic function (Montella *et al*., 2017). For example, 21 different known activities have been listed on the CAzy website for GH family 5, including endoglucanase, endoxylanase and exoglucanase. However, a large proportion of enzymes placed in GH family 5 have been annotated as having a function directly related to the breakdown of cellulose (endo- and exoglucanases), and reads assigned to this family are likely to be good candidates for carrying out an initial attack on recalcitrant polysaccharides, including crystalline cellulose. On the other hand, enzymes with varying substrate specificity and activity constitute GH families 2, 3, 9 and 10, and have the potential to mediate cellulose hydrolysis either by attacking the crystalline cellulose or intermediate products of cellulose breakdown, such as cellobiose.

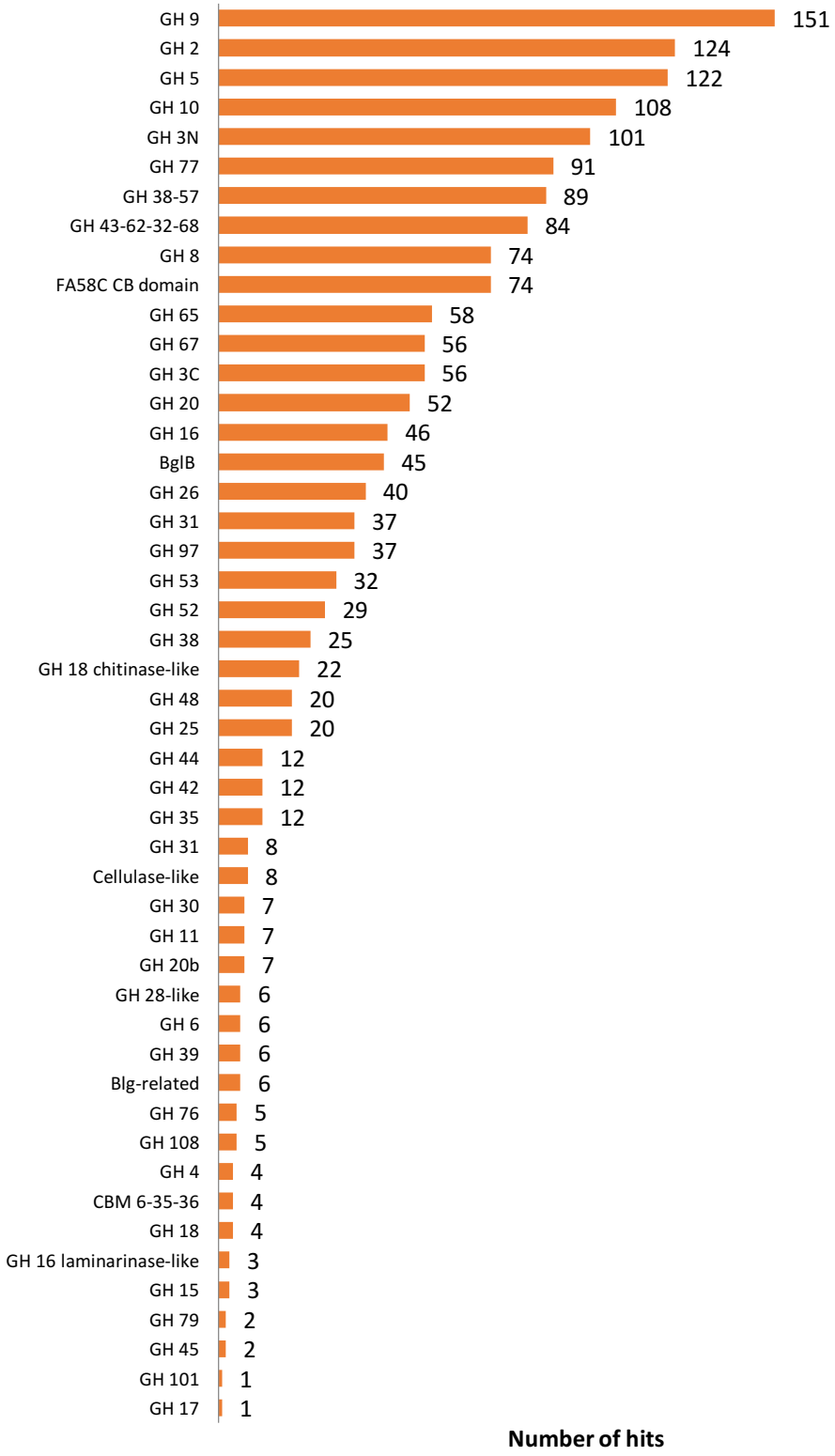


Figure 4.16 Distribution of ORFs in the metatranscriptomic data classified as having high similarity to enzymes in glycoside hydrolase families as determined by HMMer search against Pfam database.

4.3.4.2 Assessing the taxonomic origin of metatranscriptomic reads assigned to GH families

The 1,724 ORFs that were classified into GH families were also subjected to Blast analysis in order to determine the taxonomic origin of the sequences. The ORF sequences were uploaded on to Galaxy webserver www.galaxy.nbic.nl, and Blastx search was performed against the NCBI nr database using the NCBI BLAST+ tool set (Camacho *et al*., 2008). The number of search results returned for each query sequence was set to one, and an E-value threshold of 0.001 was selected. The results of the blast search for sequences with 5 or more hits are presented in Table 4.2. A total of 1,610 blast hits were returned, suggesting that 114 sequences had no matches in the NCBI nr database. These sequences potentially represent novel GH enzymes for which no homology-based close relatives could be matched in the database and could be targeted for further study, including designing of PCR primers or hybridisation probes for sequence-based screening of a fosmid library.

While the majority of the sequences only returned one hit, the highest number of hits (15) to a single ORF corresponded to a GH family 9 protein from an uncultured bacterium. Other sequences with multiple matches originated from bacteria with a reported role in lignocellulosic biomass degradation, including *Sorangium cellulosum*, *Acetivibrio cellulolyticus*, *Fibrobacter succinogenes* and *Clostridium* spp. *Sorangium cellulosum* produced multiple hits and while Genus-level relative abundance could not be determined due to poor classification using MG-RAST, Order Myxococcales which contains *Sorangium* was determined to account for 0.6% of total reads classified in the metatranscriptome. Several sequences had the blast search return closest matches to putative proteins and/or proteins originating from uncultured organisms. Moreover, the percentage identities for the matches were often low (only 42 sequences had percentage

identity >= 90%, while 398 sequences had percentage identity =< 50%), suggesting that the queried sequence might exhibit a related, yet distinct, function to the closest matched sequence in the database and that the environment under survey accommodates GH enzymes that have not been characterised previously.

Table 4.2 Results of the Blastx search of the 1,724 ORFs determined to have high similarity to glycoside hydrolases against the NCBI nr database. Columns represent the number of times a specific entry in the database was a match for an ORF query sequence, the average % identity for the matching ORFs, and annotation information.

Count	Avg. % identity	Source Organism	Annotation
15	65.9	Uncultured bacterium	Glycoside hydrolase family 9, partial
9	62.0	Sorangium cellulosum So ce56	Endo-1,4-beta-xylanase
9	79.1	Nonomuraea sp. SBT364	Beta-1 4-xylanase
9	65.4	Hahella chejuensis KCTC 2396	Endoglucanase Y
8	82.6	Acetivibrio cellulolyticus	Endoglucanase
7	63.9	Teredinibacter turnerae T7901	Glycoside hydrolase family 5 domain-containing protein
7	45.1	Fibrobacter succinogenes subsp. succinogenes S85	Glycoside hydrolase family protein
7	78.8	Acetivibrio cellulolyticus	Cellulose 1,4-beta-cellobiosidase
6	39.3	Uncultured organism	Putative carbohydrate-active enzyme
6	48.4	Ruminiclostridium thermocellum DSM 1313	Glycoside hydrolase
5	51.1	Sorangium cellulosum So ce56	Glucosylceramidase
5	49.1	Sorangium cellulosum So ce56	Glycosyl hydrolase
5	53.1	Sorangium cellulosum So ce56	Alpha-L-arabinofuranosidase
5	82.5	Clostridium cellulolyticum H10	Glycoside hydrolase
5	67.3	Anaerolinea thermophila UNI-1	4-alpha-glucanotransferase
5	47.6	Paenibacillus terrae HPL-003	Beta-xylosidase
5	86.0	Clostridium sp. BNL1100	Beta-mannanase
5	73.9	Ignavibacterium album JCM 16511	4-alpha-glucanotransferase
5	76.1	Melioribacter roseus P3M-2	Beta-hexosaminidase precursor

5	73.1	Aeromonas caviae	Xylosidase B
5	59.4	Clostridium sp. 7_2_43FAA	Glycosyl hydrolase family 20
5	49.0	Fibrisoma limi	Chitinase
5	79.3	Acetivibrio cellulolyticus	PKD domain-containing protein
5	89.6	Acetivibrio cellulolyticus	Endoglucanase
5	83.7	Acetivibrio cellulolyticus	Glycoside hydrolase
5	57.7	Saccharophagus degradans 2-40	Putative bifunctional xylanase/a-L-arabinofuranosidase

4.3.4.3 Searching the metagenome for genes encoding GHs from the metatranscriptome

We were interested in investigating the occurrence of genes that were expressed and annotated as having a function in hydrolysis of glycosidic bonds within the metagenome. The 9.07 million sequences comprising the metagenome were uploaded on to Galaxy webserver www.galaxy.nbic.nl. The previously uploaded 1,724 ORFs that were classified into GH families were used to create a custom database, and the metagenome was queried against the custom database through Blastn using the NCBI BLAST+ tool set (Camacho *et al*., 2008). Hits with 100% identity were extracted from the blast output and sequences with 50 or more hits have been presented in Table 4.3, along with annotation derived from the original metatranscriptomic read.

744 out of 1,724 sequences identified as belonging to GH families in the metatranscriptome were concluded to have exact matches in the metagenomic data, with the majority having more than 1 hit. *Acetivibrio cellulolyticus* and *Clostridium* spp. were highly represented, which is unsurprising given then Clostridiales was determined to be the most abundant Order in the taxonomic classification of the metagenome which was generated through MG-RAST analysis. The original metatranscriptomic read corresponding to some of the highly represented GH-encoding genes in the metagenome was annotated as putative or hypothetical protein in some cases. These reads are particularly interesting as they potentially represent genes in the metagenome that encode novel GHs, especially those originating from uncultured organisms. These sequences could again be targeted for further study, including designing of PCR primers or hybridisation probes for sequence-based screening of a fosmid library, especially if the starting material for the production of a fosmid library

is high molecular weight DNA extracted from the same microbial community as the metagenome under survey here (Geng *et al*., 2012).

Table 4.3 Metagenomic reads with 100% identical matches (50 or over) to genes encoding glycoside hydrolases from the metatranscriptome dataset. ID column refers to the ID assigned to the predicted reads by the MetaGeneMark ORF finder, count is the number of times a metagenomic read had a match at an identity of 100% and annotation is that originally determined for the metatranscriptomic sequence.

ID	Count	Source Organism	Annotation
gene_id_1377828	138	Clostridium clariflavum DSM 19732	Endoglucanase
gene_id_755308	127	Hahella chejuensis KCTC 2396	Endoglucanase Y
gene_id_591566	125	uncultured bacterium	Putative cellulase
gene_id_1308209	105	Clostridium josui	Hypothetical protein
gene_id_1474267	95	Acetivibrio cellulolyticus	Glycoside hydrolase family protein
gene_id_2100434	86	Clostridium clariflavum DSM 19732	Dockerin-like protein
gene_id_418817	86	Acetivibrio cellulolyticus CD2	Hypothetical protein AcelC_23359
gene_id_2257214	83	Acetivibrio cellulolyticus CD2	Hypothetical protein AcelC_15059
gene_id_1655374	68	Clostridium clariflavum DSM 19732	Beta-1,4-xylanase
gene_id_1525564	64	Clostridium clariflavum DSM 19732	Beta-glucosidase-like glycosyl hydrolase
gene_id_1896181	63	Acetivibrio cellulolyticus CD2	Family 6 carbohydrate binding protein
gene_id_1967694	62	Clostridium cellulovorans 743B	Dockerin type 1
gene_id_897903	62	Clostridium sp. Iso4-1	Glycosyl hydrolase family 48, partial
gene_id_1987598	60	Acetivibrio cellulolyticus CD2	Glycoside hydrolase family protein
gene_id_2255386	60	Thermoanaerobacterium saccharolyticum JW/SL-YS485	Glycoside hydrolase family protein
gene_id_547346	59	Acetivibrio cellulolyticus CD2	Endoglucanase
gene_id_1484959	58	Clostridium clariflavum DSM 19732	Dockerin-like protein
gene_id_560216	57	Acetivibrio cellulolyticus CD2	Glycoside hydrolase family protein
gene_id_1088883	56	Acetivibrio cellulolyticus CD2	Glycoside hydrolase
gene_id_1098361	56	Terriglobus saanensis SP1PR4	Alpha-mannosidase

gene_id_1818196	55	Clostridium clariflavum DSM 19732	Beta-mannanase
gene_id_2201343	54	Acetivibrio cellulolyticus CD2	Mannan endo-1,4-beta-mannosidase (beta-mannanase)
gene_id_1987245	53	Clostridium cellulolyticum	Cellulase Cel9-H
gene_id_1037053	51	Acetivibrio cellulolyticus CD2	Cellulase
gene_id_481913	51	Ruminiclostridium thermocellum	Endoglucanase K, partial
gene_id_577242	50	Clostridium thermocellum DSM 2360	Beta-glucosidase

4.4 Discussion

Although all the dewaxed cotton used for total nucleic acid extraction had been incubated in the same carboys, the metagenome was extracted ~ 6 months earlier than the metatranscriptome as time was required to optimise the RNA extraction protocol. Some freshly extracted mRNA was also added subsequently to the pooled aliquot of mRNA used for metatranscriptome sequencing before MessageAmp II amplification. As such, while the extracted samples can be considered parallel, there are also likely to be some differences in the microbial community colonising the cotton and comparison made between the datasets should be treated with a certain level of caution. Comprehensive comparative analyses of the three datasets can only be concluded following statistical analysis involving multiple replicates of each 'omic' dataset. Pooling of extracted samples both for the metagenomic and metatranscriptomic sequencing prevents 'omics'-mediated analysis of Bromborough Dock and Bidston Moss landfill sites. Such an analysis was never intended however, as the primary objective of the study was to identify cellulases being expressed in a landfill setting.

Illumina sequencing was chosen over another popular sequencing format, 454 pyrosequencing (now essentially obsolete), mainly due to the much greater depth of sequencing afforded with the MiSeq platform along with its relative low cost per sequence generated. While 454 pyrosequencing has been reported to generate longer read length, the output has also been prone to producing excessively short reads (< 100 bp). This issue is bypassed by MiSeq sequencing, as a more consistent read length is offered, and the paired-end sequencing chemistry allows for generation of up to a respectable 500 bp long reads (Illumina). The larger datasets produced provide their own challenges to bioinformatic analysis though, and the sheer volume of high quality

data requires not only a hefty amount of computational power, but also prolonged processing time for traditional analysis approaches such as Blast.

Despite the significantly lower error rate associated with the Illumina platform, stringent filtering and quality control of sequencing output was necessary to avoid erroneous annotations and for coherent biological conclusions to be drawn from the large datasets produced (Loman *et al*., 2012). Quality filtering steps revealed that a large proportion of metatranscriptomic reads were classed as rRNA or artificial replicate sequences. While artificial replicates are an innate feature of metatranscriptomic studies (Gomez-Alvarez *et al*., 2009), it is possible that the naturally copious amounts of rRNA present in the sample could have been artificially increased during the cDNA library preparation step before sequencing or during amplification using the MessageAmp II aRNA Amplification kit (Invitrogen).

Despite performing depletion using MICROBExpress Bacterial mRNA Enrichment Kit (Ambion) and Terminator 5' Exonuclease (Epicentre), SSU and LSU rRNA accounted for a significant proportion of the metatranscriptomic sequence reads (>70%). Stewart *et al*. (2010) found that rRNA made up 51-60% of the total RNA sequences in their metatranscriptomic dataset, only subtractive hybridisation was used in this study and not the subsequent enzymatic depletion. In the study, they generated custom hybridisation probes tailored to the microbial community under investigation, an approach that can lead to immensely improved rRNA removal. Such an effort, however, will prove to be problematic for environments that consist of microbes that have not been well characterised. Other studies have reported the proportion of rRNA reads in metatranscriptomic datasets within the range of 37%-90% (Gifford *et al*., 2011; Shi *et al*., 2011, de Menezes *et al*., 2012). Therefore, it is not impossible that some rare

transcripts were not sequenced and were consequently omitted from the analyses, despite given the greater sequencing depth afforded with Illumina MiSeq sequencing.

Lack of annotated sequences in the database, the inefficiency of designed probes and the targeting of small amounts of 5'monophosphorylated mRNA by Terminator 5' Exonuclease pose further complications (Mettel *et al*., 2010). While eukaryotes possess naturally polyadenylated mRNA transcripts that aid their separation from rRNA sequences, archaeal and bacterial mRNA transcripts have no poly-A tails. He *et al*. (2010) assessed rRNA removal as well as the effect of these procedures on the composition and integrity of the community mRNA and found subtractive hybridisation to be an effective method for rRNA removal, particularly as it had no effect on the mRNA profile. On the contrary, exonuclease treatment was reported to target partially degraded mRNA molecules lacking a 5'triphosphate. However, it is worth bearing in mind that the subject of this study was an artificial community composed of 5 defined microbial culture isolates, and that a naturally occurring microbial community will harbour microbes that are not present in the database, leading to insufficient subtractive hybridisation using a commercial kit. Abundance of mRNA has been documented to vary between different microbial species and it has been estimated to make up only between 1-10% of the total community RNA (He *et al*., 2010). As such, rRNA content of ~72% in the metatranscriptome is a significant improvement on the amount of rRNA reported in non-depleted metatranscriptomes.

Following quality filtering, millions of high quality metagenomic and metatranscriptomic reads were profiled for taxonomic as well as functional annotations. While MG-RAST generates a broad overview of the functional make-up of the microbial community, it is important that such classifications are viewed as a best estimate result and not as conclusive proof for the presence or absence of enzymes

mediating key activities in the datasets. Nonetheless, MG-RAST is a valuable resource for high throughput sequence analysis as its workstation option allows for in-depth analysis of reads assigned through specific databases. This is particularly true for functional annotations, as the user is able to focus specifically on reads classified using KEGG, COG, SEED subsystems as well as NOG databases.

It was notable that only a small proportion of metatranscriptomic reads could be assigned to a functional category with a relatively stringent E-value threshold of 0.001, as only 167,005, 115,113 and 300,798 reads could be classified using COG, KEGG and SEED subsystems, respectively. A large number of reads were assigned to functional categories involved in maintaining general 'housekeeping' functions of the microbes, while detailed analysis of some reads suggested potential function in cellulose hydrolysis. For example, it was evident that carbohydrate metabolism was central to the functional profile of the microbial community, and further analysis of this category highlighted hits documented as glycoside hydrolases, endoglucanases and structural components of cellulosomes. These functional annotations back the taxonomic analysis suggesting that cellulolytic microbes are present and active in the colonised cotton incubated in landfill leachate.

It was determined that a higher percentage of reads in the amplified metatranscriptome corresponded to rRNA and artificial replicate compared to the non-amplified metatranscriptome. This is possibly an artefact of the amplification step in the MessageAmp II kit as enzymatic steps are required for dscDNA generation, and could have led to preferential amplification of the already ample rRNA sequences present in the sample. While the majority of the comparative analysis between the two metatranscriptomes generated similar results, some discrepancies were also observed. For example, the percentage of Archaea in the amplified metatranscriptome were lower

compared to the metatranscriptome, while the percentage of reads assigned as Proteobacteria were higher and the percentage of Unclassified (derived from Bacteria) were lower in the phylum level taxonomic analysis. Functional annotations were largely determined to be similar. Analysis of several replicates is required to conclude if there is a significant difference in the community profiles between the two metatranscriptomes, as small differences seen here could be due to temporal variations between the singular samples analysed. Feldman *et al*. (2002), Polacek *et al*. (2003) and Li *et al*. (2004) reported that any bias introduced as a result of mRNA amplification using the MessageAmp II aRNA amplification kit was minimal.

Blast analysis of high throughput sequencing datasets frequently reveals closest matches of mRNA transcripts to be hypothetical/predicted proteins, often originating from unclassified/uncultured microbes. Assigning reads to families of proteins using HMMer algorithm to search against the curated Pfam database is a more informative strategy and provides more precise data on function of the sequences queried. However, since searching against HMM models is a more sensitive database search than Blast, it is also slower. The assignment of 1,724 predicted ORFs to GH families followed by a Blast search against NCBI nr database revealed some sequences to have no match in the database suggesting aspects of novelty, perhaps based on structure or catalytic activity, and could be applicable in industry based on those characteristics. Moreover, metatranscriptomic sequences assigned to GH families were determined to be present in metagenomic data, strongly suggesting that genes represented in these sequences are present in detectable amounts in the environment. The fact that both the metagenome and the metatranscriptome are derived from microbes colonising dewaxed cotton hints towards a defined role for the enzymes in cellulose decomposition.

There are a number of issues and difficulties that influence the generation and analysis of metagenomic and metatranscriptomic datasets. The isolation of good quality nucleic acids as starting material is the single most important factor influencing the integrity and completeness of data obtained from any 'omics' study (Engelbrektson *et al*., 2010). The stability of mRNA samples is one of the key limitations applicable to metatranscriptomics studies, with the average half-life of mRNA molecules reported to be in the range of minutes to seconds (Deutscher, 2006; Presnyak *et al*., 2015), whilst also being under the influence of the nutritional status of individual cells (Redon *et al*., 2005). The rate of mRNA degradation is variable between different microbial species and is noted as being similar for genes that share biological functions (Bernstein *et al*., 2002; Selinger *et al*., 2003; Hambraeus *et al*., 2003).

Adsorption of RNA molecules to solid matter in the leachate is another issue, particularly as the rate of adsorption increases in low pH and high salt, conditions that are typical for RNA extraction buffers (Chomczynski & Sacchi, 1987). The quality of the extracted nucleic acids can also be compromised due to the presence of compounds that either interfere with the extraction protocol, or inhibit downstream molecular biological applications. These include DNase and RNase enzymes synthesised by a highly diverse microbial contingent, as well as the abundance of complex organic molecules including humic and fulvic acids found inherently associated with the leachate (Barlaz & Ham, 1993).

Bioinformatic analysis of sequence data derived from an individual microbe isolated in culture is relatively simple, particularly if a reference genome is available. However, given the highly diverse and microbially populous nature of environmental samples such as landfill leachate (Thomas *et al*., 2012), coupled with the huge amount of heterogeneity it displays, analysis of metagenomic and metatranscriptomic data

remains computationally demanding (Prakash & Taylor, 2012; Aguiar-Pulido *et al*., 2016).

Despite the recent advances in sequencing technologies, the goal to completely reconstruct genomes belonging to discrete microbes within a complex environmental sample remains largely unattainable due to technological limitations in sequencing throughput and bioinformatic analyses (Scholz *et al.*, 2012; Abram, 2015). The length of sequence reads has improved but they still remain relatively small, meaning crippling amounts of computational prowess is required for assembly of overlapping reads. Similarity of closely related sequences leading to chimeric assembly and the low abundance of rare genomes causing poor-to-no assembly at all are the other major issues facing assembly of complete genomes within metagenomes (Xie *et al*., 2010).

Alignment of unassembled reads against a database has been commonly applied to metagenomic datasets when assembly is either not possible or incomplete. However, Basic Local Alignment Search Tool (BLAST) algorithm is unsuited for aligning millions of sequence output reads from a shotgun metagenomic run on the Illumina platform to a large database such as NCBI nr protein database. Taxonomic and functional annotations of metagenomic data can be visualised innovatively using MEGAN software (Huson & Mitra, 2012). However, data analysis is complicated due to MEGAN's requirement for a Blastx output file to visually summarise the annotations. MG-RAST (Meyer *et al*., 2008) is a web service designed specifically for the functional and taxonomic analysis of metagenomic and metatranscriptomic data as it utilises algorithms that are more efficient than Blast. However, the inability to customise the analysis pipeline for specific datasets and the growing popularity of stand-alone software is beginning to cause issues, particularly with the increase in the volume of data being generated per sequencing run. Webserver-based alternatives like

IMG/M (Markovitz *et al*., 2014), WebMGA (Wu *et al*., 2011), METAREP (Goll *et al*., 2010) and COMAN (Ni *et al*., 2016) also allow for taxonomic and functional profiling of metagenomic and metatranscriptomic reads from microbial communities.

Development of software such as Metabin (Sharma *et al*., 2012) and GenoMeta (Davenport *et al*., 2012) aims to circumvent issues with the alignment of short metagenomic reads with a high degree of confidence. While GenoMeta employs the Bowtie alignment tool to generate a histogram of read distribution, Metabin utilises Blast-Like Alignment Tool (BLAT) (Kent, 2002) for relatively quick generation of output that can be visualised in various styles. MetaBAT (Kang *et al*., 2015) and MetaCluster-TA (Wang *et al*., 2014) are more recent tools developed for taxonomic binning specifically designed with metagenomic data in mind. However, none of these enable functional annotation of metagenomic reads for characterisation of potentially novel genes of interest, as they focus solely on taxonomic classification.

Only a subset of the sequences obtained from high throughput sequencing of environmental samples can be functionally or taxonomically classified owing to limitations in the size and quality of metagenomic databases. While short read aligners can align vast amount of reads from a single organism to a reference sequence efficiently, mapping sequenced reads is not always possible due to the lack of relevant reference genomes in the databases. This issue can now be circumvented, as software for the *de novo* assembly of genomic reads exist. However, these programs are computationally demanding, and the problem is exacerbated when the data involved is metagenomic and metatranscriptomic. Output data from high throughput sequencing is characterised in terms of depth and coverage. The absolute number of reads generated is referred to as depth, while the number of times each base is called is referred to as coverage. High coverage is ideal to facilitate assembly or to allow analysis of Single

Nucleotide Polymorphisms (SNPs). Different samples producing the same depth of sequencing will generate different levels of coverage. While it is possible to achieve coverage of 5-10 fold when sequencing single genomes, metagenomes and metatranscriptomes only get fractional coverage making accurate assemblies practically impossible. Draft genome assembly is usually incomplete, and generating the level of coverage necessary for complete assembly is prohibitively expensive (Hess *et al*., 2011). Metagenomic shotgun reads usually consist of viral as well as eukaryotic sequences, along with bacterial and archaeal, and there is currently no centralised system available that combines all necessary pipelines for the analysis of such complex samples (Scholz *et al.*, 2012; Abram, 2015).

4.5 Conclusions

This chapter outlines the generation and bioinformatic analysis of high throughput 'omics' data derived from an environmental mixed microbial community. Taxonomic and functional analyses of a metagenome, a metatranscriptome and an amplified metatranscriptome were undertaken, and the major outcomes are outlined below:

- Quality control and pre-processing of the data was performed using Galaxy (Afgan *et al*., 2016, accessible at www.usegalaxy.org) for the removal of short and low quality sequences, as well as artificial duplicate sequences. Paired-end reads were combined to generate a total of 9.07 million, 9.59 million and 8.58 million sequences from the metagenome, the metatranscriptome and the amplified metatranscriptome, respectively.
- Domain-level taxonomic classification performed using MG-RAST (Meyer *et al*., 2008) suggested that the three datasets were dominated by Bacteria.

Percentage of reads assigned to Archaea increased from 6.6% to 14.2%, and those assigned to Eukaryota increased from 1% to 7.6% in the metatranscriptome compared to the metagenome; suggesting that members representing these domains are highly active in the environment under survey.

- Proteobacteria, Firmicutes, Bacteroidetes, Planctomycetes and Actinobacteria contained the most bacterial representatives in the metagenome and the metatranscriptome, while Euryarchaeota was the most populous archaeal phylum. Unclassified sequences comprised ~ 11% of total microbial sequences in the metatranscriptome compared to only 1.2% in the metagenome, suggesting that a significant number of never sequenced before microbes are active in the environment.
- Percentage of unclassified reads from domain Eukaryota increased from ~ 14% in the metagenome to ~ 23% in the metatranscriptome; while Family Metopidae, which includes genera with a reported role in cellulose degradation such as *Metopus* and *Nyctotherus*, accounted for 1% of all Eukaryota reads and 7% of all unclassified reads derived from Eukaryota in the metatranscriptomic data.
- Assignment of annotated proteins to SEED subsystems (Overbeek *et al*., 2005) categories suggested that the datasets are dominated by reads with a function in Clustering-based subsystems, Carbohydrate metabolism, Protein metabolism, RNA metabolism and processing of amino acids and derivatives, pointing towards an active microbial community.
- Attempts made at determining taxonomic classification using MetaPhlAn (Segata *et al*., 2012) and the mini version of Kraken (Wood & Salzberg, 2014) were unsuccessful, as the overwhelming majority of reads could not be

classified. This suggests that tools that rely on marker genes and k-mers derived from reference genomic libraries are perhaps not suitable for taxonomic classification of complex environmental microbial communities.

- MetaGeneMark (Zhu *et al*., 2010) was used to predict 2.34 million ORFs in the metatranscriptomic mRNA reads, and a total of 1,724 ORFs were determined to have a high similarity to a glycoside hydrolase family protein in the Pfam-A protein database. The highest number of hits were assigned to GH families 9, 2, 5, 10 and 3.
- The taxonomic origin of the 1,724 ORFs was determined using Blastx, with 114 sequences returning no matches in the NCBI nr database and hence representing potentially novel GH enzymes. Bacteria with a reported role in lignocellulosic biomass degradation, including *Sorangium cellulosum*, *Acetivibrio cellulolyticus*, *Fibrobacter succinogenes* and *Clostridium* spp., returned multiple hits, while the highest number of hits (15) to a single ORF corresponded to a GH family 9 protein from an uncultured bacterium.
- Of the 1,724 ORFs identified as belonging to a GH family in the metatranscriptome, 744 were found to have exact matches in the metagenomic data, with proteins originating from *Acetivibrio cellulolyticus* and *Clostridium* spp. highly represented.

Chapter 5

Production of fosmid libraries from environmental DNA and screening for lignocellulolytic enzymes

5.1 Background

Gene mining for novel enzymes that have industrial and biotechnological applications from understudied environments harbouring complex and robust microbial communities is a key area of research in environmental metagenomics. A functional metagenomics approach was applied here, whereby the community DNA from anoxic lake sediment and from anoxic landfill leachate was cloned and expressed in a heterologous host for the discovery of novel lignocellulosic enzymes. Whilst BACs can incorporate DNA of length over 100 kb, fosmids were chosen as the vectors due to the relative ease of working with the latter coupled with the fact that high molecular weight (HMW) DNA between the size 30-50 kb is still likely to contain whole genes and even gene clusters consisting of ancillary and regulatory elements for the purpose of efficient screening. This is particularly applicable to cellulases from anoxic environments, as specific secretion pathways as well as assemblages of structural proteins that form the cellulosomes may be necessary. Therefore, it was hypothesized that cloning and screening of HMW DNA extracted directly from microbial community resident in anoxic environments could lead to the successful isolation of potentially novel glycoside hydrolases.

Traditionally, the hit rates for the discovery and isolation of glycoside hydrolases have been reported to be very low from environmental metagenomes (Kakirde *et al*., 2010). Dewaxed cotton cellulose bait was used in this study to enrich

for cellulose-degrading microbes in an attempt to improve the hit rate for novel glycoside hydrolases. Such a method has been shown to increase the yield of cellulases in a pyrosequenced metagenome (Edwards *et al*., 2010). A major advantage conferred by the functional screening of shotgun metagenomic libraries is that it can lead to the isolation of an actual gene product, rather than just *in silico* data in the form of sequence reads. Genes for enzymes of interest can then be sub-cloned, overexpressed and characterised with a view to biotechnological applications in industry. Such an approach has led to the identification of cellulases from termite gut (Warnecke *et al.*, 2007; Nimchua *et al*., 2012), buffalo rumen (Nguyen *et al*., 2012), grassland soil (Nacke *et al*., 2012) and Sargasso Sea (Cottrell *et al*., 2005), amongst other environments, but not yet from landfill or freshwater lakes.

One of the most common methods for the functional screening of fosmid libraries for the isolation of endoglucanases is the Congo red assay (Teather & Wood, 1982). The assay incorporates carboxymethylcellulose (CMC), a soluble derivative of cellulose, into agar plates that are subsequently stained with Conge red dye. The dye reacts with CMC to present the plates with a red appearance, and the clones expressing endoglucanases appear to have a zone of clearance around them as a result of the enzymatic hydrolysis of CMC (Figure 5.1). Using this method, cellulases have been successfully recovered from soil (Voget *et al*., 2006), buffalo rumen (Duan *et al*., 2009) and the intestinal tract of abalone (Kim *et al*., 2011) and termite (Zhang *et al*., 2011), amongst others. Assays incorporating other substrates have also been described for the detection of endoglucanases and endoxylanases. Broth cultures of clones can be assayed for cellulase activity using the *p*-nitrophenyl β-D-cellobioside (pNPC) assay (Deshpande *et al*., 1988), a method that has been used to compare the cellulolytic activity of actinomycetes (Ball *et al*., 1992). The enzymatic breakdown of pNPC

releases *p*-nitrophenol (pNP), levels of which can be measured using spectrophotometry.

Furthermore, AZCL-HE-Cellulose and AZCL-Xylan from birchwood (Megazyme, Ireland) are substrates generated by crosslinking azure dye with insoluble cellulose and xylan, respectively. These substrates can be used to screen clones either in broth culture (Baldrian *et al*., 2005) or on agar plates that have been overlaid with top agar containing the appropriate substrate (Nguyen *et al*., 2012). The hydrolytic action of the enzymes cleaves the bond between the cellulose or the xylan and the blue dye, leading to its release either in the assay mixture or in the top agar. Clones expressing enzymes of interest can be detected in broth-based assays by spectrophotometry and those in agar-based assays can be detected visually as they appear to have a blue zone around the colonies.

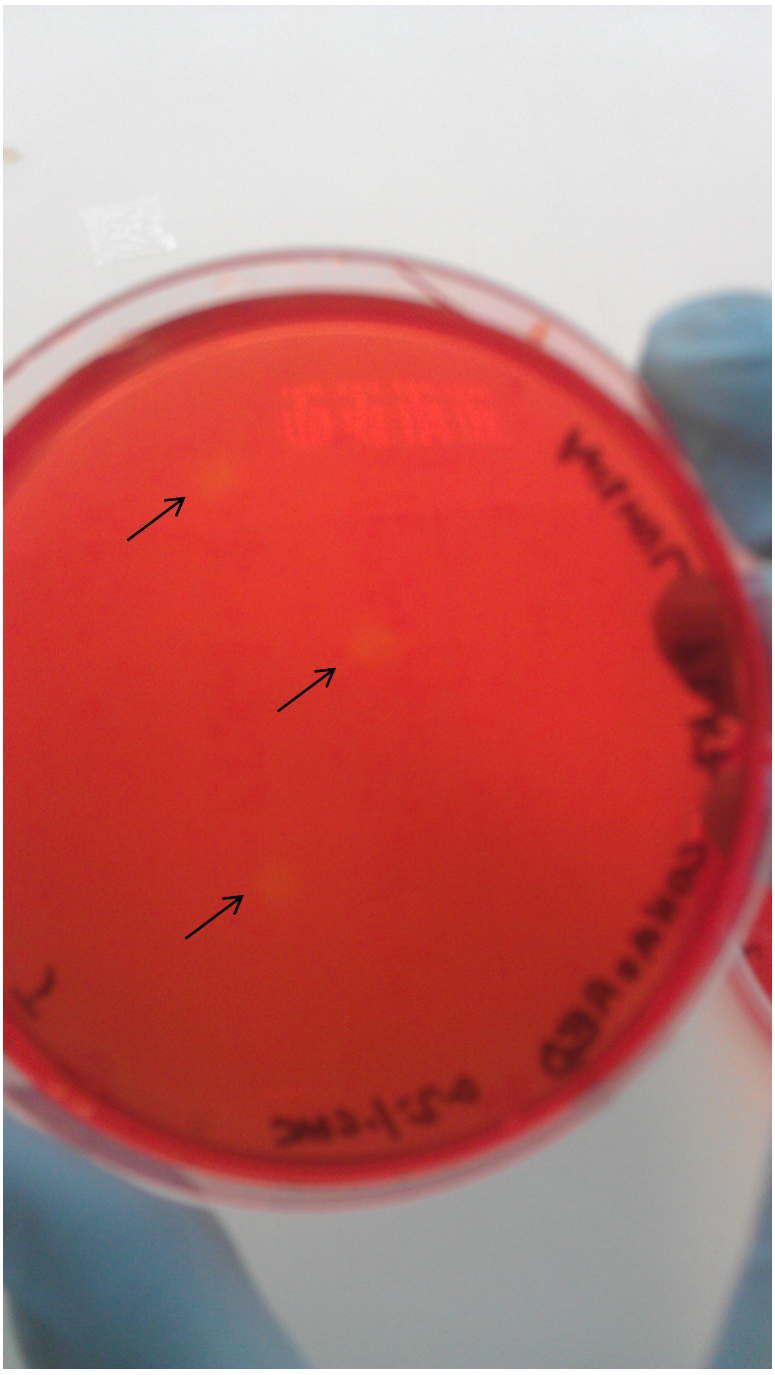


Figure 5.1 Example of a positive Congo red reaction from this study. 3 zones of clearance can be seen on the plate against a red background.

5.2 Extraction of HMW DNA for the production of a fosmid library

To generate a fosmid library for the screening and isolation of lignocellulosic enzymes originating from landfill, HMW DNA was extracted and pooled from samples in parallel to those used to generate metagenomic and metatranscriptomic data (see chapter 4), i.e. colonised string incubated in leachate carboys BD1, BD2, BD3, BM3E, BM3F, BM3G and BM1J. Although the DNA extracted using the method described by Griffiths *et al*. (2000) (section 2.2.2) is perfectly suitable for downstream processes such as PCR amplification and high throughput sequencing due to the high level of purity observed, its size range is typically below 11 kb. This is due to the protocol involving mechanical disruption of the cells through high-speed bead beating. As such, the DNA tends to be highly sheared for use in the production of a fosmid library, and a much gentler method was required to produce the HMW DNA. Three extraction protocols were tested to determine their suitability for extracting DNA within the size range 30 kb-50 kb:

- High molecular weight DNA extraction method (Neufeld *et al*., 2007)
- High molecular weight DNA extraction method modified from Neufeld *et al*. (2007)
- Meta-G-Nome DNA Isolation Kit (Epicentre)

5.2.1 High molecular weight DNA extraction method (Neufeld *et al*., 2007)

This method (section 2.2.4) relies on enzymatic lysis of cells and consists of gentle washing steps whilst still incorporating phenol:chloroform (24:1) extraction for sufficient purity of DNA to be attained. Another advantage is the use of phase-lock gel (PLG) tubes that separate the top aqueous layer from the bottom organic layer containing various dissolved impurities by forming a stratified barrier between them

after centrifugation. Therefore, the aqueous layer consisting of the nucleic acids can be decanted off without the requirement for extra pipetting, as multiple pipetting steps can potentially shear the DNA. The sample was not vortex mixed at any stage for a similar reason.

The extraction protocol was tested on half a piece of string (~ 1 g wet weight) incubated in carboy BD3. A total of 400 ng of DNA was extracted. However, its purity was determined to be insufficient for the generation of a fosmid library as the A_{260}/A_{280} absorbance reading from the NanoDrop 2000 (Thermo Scientific) was only 1.59 (Table 5.1), suggesting sample contamination by either protein or phenol. As an A_{260}/A_{280} ratio of >1.70 is necessary, the extracted DNA was cleaned up using the UltraClean Soil DNA Isolation Kit (MO BIO) by following only the spin column-based steps, according to the manufacturer's instructions. As expected, the processed sample was assessed to be extremely clean, with an A_{260}/A_{280} ratio of 2.23. However, a common undesirable consequence of using spin column-based methods for the clean-up of nucleic acids is the loss of yield, as <250 ng of DNA was recovered. Since the source material of the colonised string was finite, the observed yield was deemed to be too low when half a string was sacrificed, and extractions using this method were not pursued further.

5.2.2 High molecular weight DNA extraction method modified from Neufeld *et al*. (2007)

Dr. James Houghton (University of Liverpool) modified the extraction protocol described by Neufeld *et al*. (2007) to extract HMW DNA from colonised cotton incubated in the anoxic sediment of freshwater lakes. The main modifications included the omission of sucrose from the lysis buffer, and the inclusion of a phenol:chloroform:isoamyl alcohol (25:24:1) extraction step to replace the use of PLG

tubes (section 2.2.5). The latter modification in particular was introduced to try and limit the excessive protein and/or phenol carryover in the extracted DNA. The use of CTAB buffer was retained to remove the excessive humic and fulvic acids usually found associated with samples originating from landfill leachate. To prevent mechanical shearing of the DNA, care was taken not to vortex mix the sample at any stage and the number of pipetting and centrifugation steps were limited wherever possible.

The extraction protocol was tested on one whole string (~ 2 g wet weight) incubated in carboy BD3. A total of 12.10 μg of DNA was extracted with an A_{260}/A_{280} reading of 1.78 (Table 5.1). Since a large amount of DNA of sufficient purity was extracted, pulsed field gel electrophoresis (PFGE) was performed using a 1% (w/v) agarose gel (section 2.9.1) to determine the fragment size of the DNA (Figure 5.2a). A large smear of DNA was observed, with the size ranging from >50 kb to <10 kb. Despite the presence of a 'tail' of sheared DNA <25 kb, the method was deemed to be suitable for HMW DNA isolation for the preparation of a fosmid library as a significant proportion of the DNA was examined to be within the size range 30 kb-50 kb and could be separated from the rest of the unsuitable DNA by size fractionation. Therefore, DNA was extracted using this method from one string each incubated in carboys BD1, BD2, BM3E and BM1J before being pooled with the extract from BD3, as those were the only carboys at that stage that consisted of incubated string demonstrating microbial colonisation and cellulose degradation. The average A_{260}/A_{280} ratio was determined to be 1.80 (Table 5.1) and a total of 55.90 μg of DNA was obtained.

5.2.3 Meta-G-Nome DNA Isolation Kit (Epicentre)

Epicentre describes the Meta-G-Nome DNA Isolation Kit as capable of extracting DNA from uncultivable environmental microbes from water, soil or compost. The kit utilises filtration and enzymatic lysis to extract DNA free from humic and fulvic acids, despite the exclusion of phenol and CTAB buffer. The manufacturers also claim that randomly sheared HMW DNA can be efficiently extracted for direct application in fosmid cloning without the requirement for size fractionation. HMW DNA was extracted according to the manufacturer's protocol (section 2.2.6), with the following modifications: the use of sterile Miracloth filtration material and 1.2 μm filter for pre-filtering the sample was excluded, and the cell pellet was resuspended in a combination of TE buffer and 0.33 volume 5% CTAB buffer before the start of the lysis protocol. The latter step was introduced due to the exceptionally humic nature of the sample. As previously, care was taken not to vortex mix the sample at any stage and the number of pipetting and centrifugation steps were limited wherever possible to prevent mechanical shearing of the extracted DNA.

Table 5.1 Table demonstrating the difference in DNA yield and purity (assessed as the A_{260}/A_{280} ratio) observed during the extraction of HMW DNA from colonised cellulose string BD3. Total DNA yield and average purity for methods deemed to be successful in extracting HMW DNA for the production of a fosmid library are also reported.

Extraction method	**Weight of string used (g)**	**DNA yield (μg)**	**A_{260}/A_{280} ratio for string**	**Total DNA yield from all strings (μg)**	**Average A_{260}/A_{280} ratio for all strings**
Neufeld *et al.* (2007)	1.0	0.4	1.59	-	-
Modified Neufeld *et al.* (2007)	2.0	12.1	1.78	55.9 *	1.80 *
Meta-G-Nome DNA extraction kit (Epicentre)	0.6	3.0	1.66	14.5 **	1.72 **

* Extracted from strings BD1, BD2, BD3, BM3E and BM1J
** Extracted from strings BD1, BD2, BD3, BM3E, BM1J, BM3F and BM3G

The extraction protocol was tested on one third of a string (~ 0.6 g wet weight) incubated in carboy BD3. A total of 3.0 μg of DNA was extracted with an A_{260}/A_{280} ratio of 1.66 (Table 5.1). Even though the purity was assessed to be slightly lower than ideal, PFGE was performed using a 1% (w/v) agarose gel (section 2.9.1) to determine the fragment size of the DNA (Figure 5.2a). As the kit suggested, a prominent band of DNA was visible at ~ 40 kb with very little sheared DNA visible as a 'tail' <40 kb. Further HMW DNA was extracted using this method from string incubated in carboys BD1, BD2, BM3E, BM1J, BM3F and BM3G before being pooled with that from BD3. The average A_{260}/A_{280} ratio was determined to be 1.72 (Table 5.1) and a total of 14.50 μg of DNA was extracted. Given the sufficient purity observed from an average extract and the excellent size profile of the DNA, the method was conceived to be suitable for HMW DNA extraction for the subsequent generation of a fosmid library.

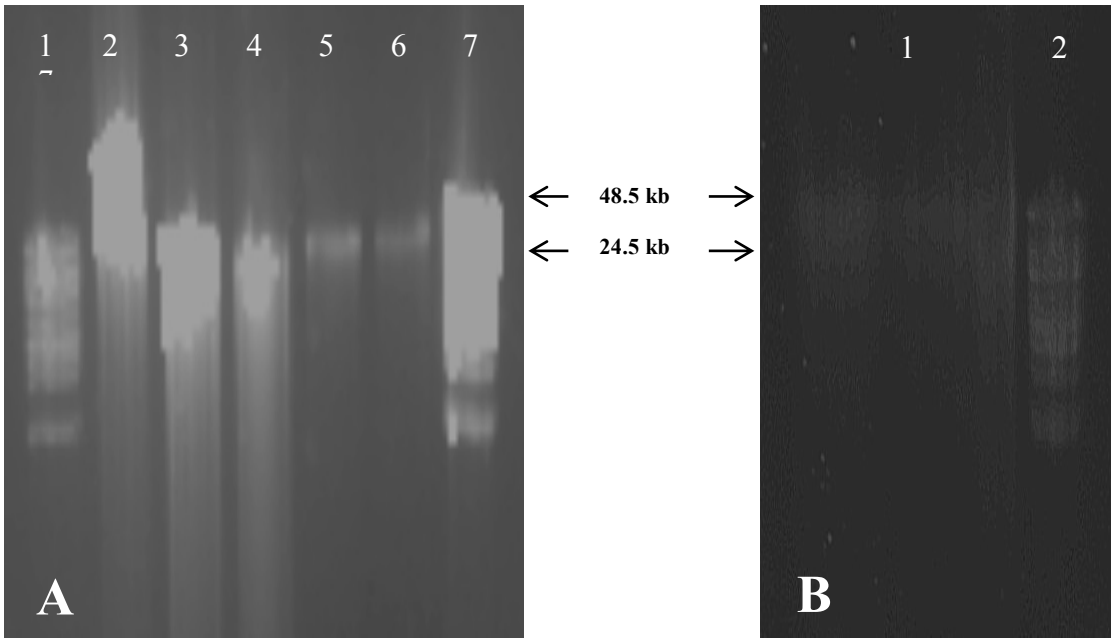


Figure 5.2 Pulsed field gel electrophoresis of a selection of extracted HMW DNA samples. (A) Lanes 1 and 7, size marker (GeneRuler High Range DNA Ladder, Thermo Scientific); lane 2, DNA extracted from string BD3 using modified Neufeld *et al*., (2007) method; lane 3, DNA extracted from string BM3E using modified Neufeld *et al*., (2007) method; lane 4, DNA extracted from string BM1J using modified Neufeld *et al*., (2007) method; lane 5, DNA extracted from string BD3 using Meta-G-Nome DNA extraction kit (Epicentre); lane 6, DNA extracted from string BM3G using Meta-G-Nome DNA extraction kit (Epicentre). (B) Lane 1, combined HMW DNA that was used for the production of the landfill fosmid library, extracted from strings BD1, BD2, BD3, BM3E and BM1J using the modified Neufeld *et al*., (2007) method, and from strings BD1, BD2, BD3, BM3E, BM1J, BM3F and BM3G using the Meta-G-Nome DNA extraction kit (Epicentre); lane 2, size marker (GeneRuler High Range DNA Ladder, Thermo Scientific).

5.3 Production, replication and storage of fosmid libraries

Total HMW DNA extracted using the modified Neufeld *et al*. (2007) method and the Meta-G-Nome DNA Isolation Kit (Epicentre) from string incubated in landfill leachate was pooled (Figure 5.2b) before being used for the generation of a fosmid library using the CopyControl^{TM} Fosmid Library Production Kit (Epicentre) following the manufacturer's protocol. The use of CopyControl^{TM} Fosmid Library Production Kit allows for regulation of the vector copy number, as the fosmid vector contains both the single-copy F-plasmid origin of replication as well as the *ori*V origin of replication to allow high-copy plasmid replication. Addition of L-arabinose or AutoInduction solution (Epicentre) induces the promoter for the *trf*A gene within the EPI300^{TM}-T1^{R} *E. coli* strain provided with the kit, allowing the mutant *trf*A gene product to initiate replication from *ori*V. Maintaining single copy during culturing and storage ensures stability of the clones, whereas up to 100 copies of the vector can be induced per cell prior to screening. This is particularly useful as single-copy conditions are prone to yield false-negative results during assays (Martinez *et al*., 2007; Martinez *et al*., 2010). The fosmid vector incorporates a gene for chloramphenicol resistance, which serves as the antibiotic selection marker.

Firstly, ~ 5 μg of metagenomic DNA was end-repaired to generate blunt-ended, 5'-phosphorylated DNA. This complements the cloning-ready, linearised and dephosphorylated fosmid vector during ligation, while the dephosphorylation prevents the self-ligation and re-ligation of the vector. Due to the wide size range of the extracted DNA, it was size fractionated by performing a PFGE using 1% low melting point (LMP) agarose (section 2.5). It was necessary to prevent ethidium bromide contamination and UV light-mediated damage to the DNA prior to cloning. As suggested in the kit instructions, control DNA and molecular size markers were run in

the outside lanes of the gel and the end-repaired DNA run in the middle lanes (Fig. 5.3). This facilitated excision of the size marker lanes exclusive of the end-repaired DNA to be stained using ethidium bromide followed by UV visualisation and marking the area of the gel containing DNA of size 30 kb- 50 kb. The stained marker sections were subsequently lined up with the unstained part of the gel, and sections containing HMW DNA of the desired size were excised to yield ~ 550 ng of DNA, ~ 200 ng of which was used for the ligation reaction. The ligated DNA was packaged into lambda phage, which was subsequently used to infect an *E. coli* EPI300-T1^{R} plating strain and selected on LB agar plates containing 12.5 μg/ml chloramphenicol. A total of 88 plates consisted of ~ 1085 colonies each, yielding a representative library of ~ 95,000 clones.

A further attempt was made at generating more clones. Surprisingly however, no clones were obtained. Titre determination of the control DNA run alongside the environmental DNA suggested that cloning efficiency declined by ~ 160-fold, as the titre decreased from 3.2×10^{7} cfu ml^{-1} to 2.0×10^{5} cfu ml$^{-1.}$ This may be attributed to the dated nature of the kit used for cloning, as it was found to be a couple of months past its expiry date. However, since the kit was only used 3 times when purchased and had since been stored frozen at -20ºC for over 12 months, it is stipulated that it served the purpose during the first attempt but not following subsequent freeze-thaw cycles.

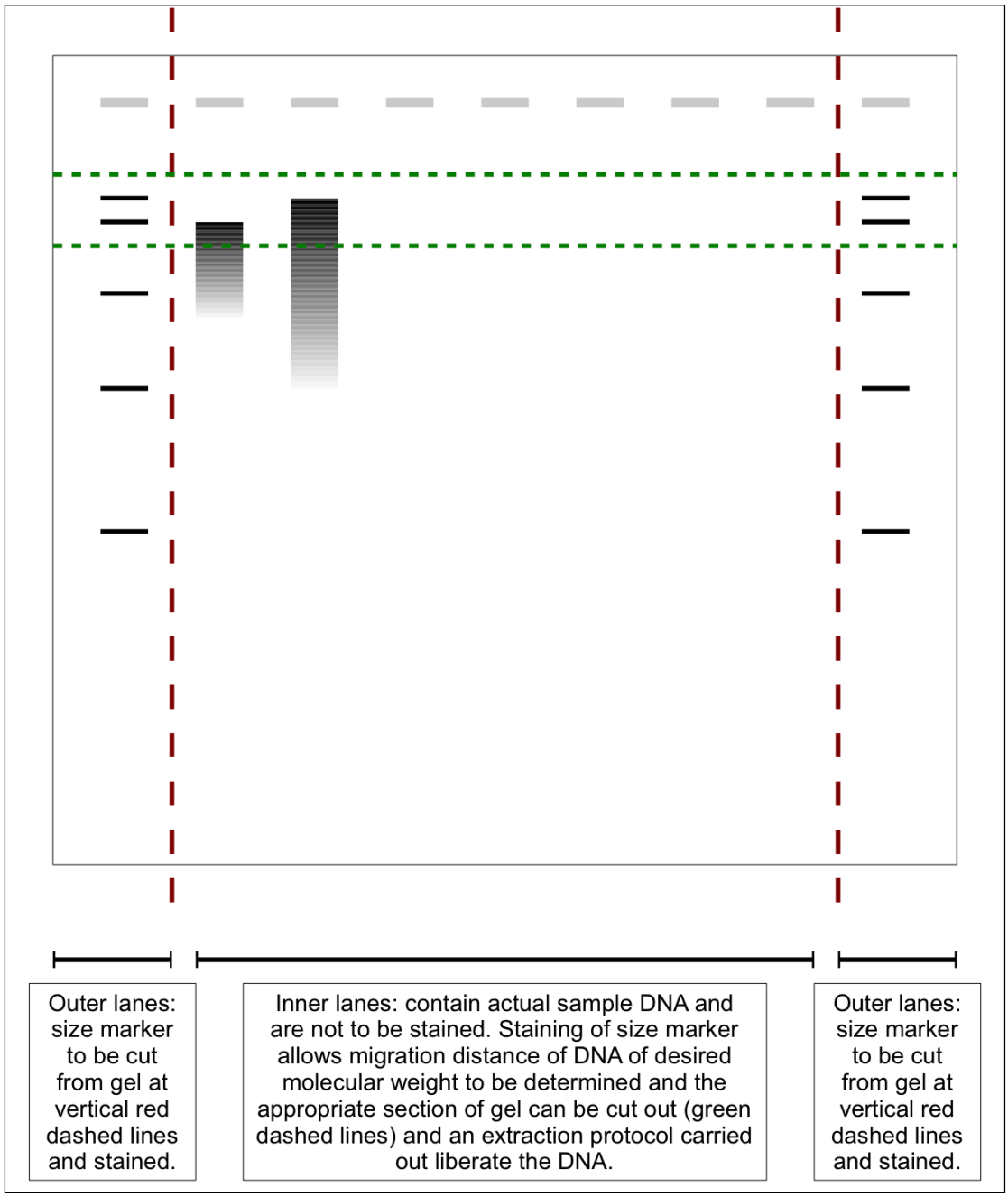


Figure 5.3 Methodology for the excision and liberation of size selected HMW DNA from PFGE gels within the size range 30-50 kb.

The fosmid clones were picked and propagated in freezing medium in 2 ml MasterBlock 96-well plates (Greiner), with each deep well consisting of a pool of ~ 500 fosmid clones. The library was maintained in this manner due to logistic and practical considerations when replicating and storing the library. A second fosmid library was produced by Bio S&T (Montreal, Canada) incorporating HMW DNA extracted by Dr. James Houghton from colonised cotton that was incubated in the anoxic sediment of freshwater lake Esthwaite Water (English Lake District). The commercial company also used the CopyControl^{TM} Fosmid Library Production Kit and delivered ~ 80,000 clones in a similar 96-well plate format, with each deep well consisting of a pool of 400-500 clones. The fosmid libraries were replicated into further 2 ml 96-well plates and stored at -80ºC until subsequent use for screening.

5.4 Functional screening of the lake-derived fosmid library for isolation of lignocellulosic enzymes

Functional screening of HMW DNA metagenomic libraries provides a relatively straightforward way of detecting the expression of proteins of interest, and is advantageous over sequence-based screening methods as no prior knowledge of the target genes is required. The lake-derived fosmid library was screened using pNPC assay, AZCL-HE-Cellulose assay, AZCL-Xylan assay and Congo red assay as described in the following sections. All assays were performed using commercially available cellulase or xylanase (Sigma-Aldrich) as positive control, and a culture of *E. coli* DH5α as negative control.

5.4.1 AZCL-Xylan assay

The broth-based AZCL-Xylan assay was performed in 96-well format to investigate which pools of 400-500 clones contained individual clones positive for expression of endoxylanases, as described in (section 2.6.3). The pools of clones were induced to high fosmid copy number to aid detection using AutoInduction solution (Epicentre) following the manufacturer's protocol. Pool H5 in plate 1 and pools B5 and H2 in plate 2 displayed significantly higher absorbance (Table 5.2).

5.4.2 pNPC assay

The pNPC assay was performed on the lake-derived fosmid library following induction of high copy number of fosmids in pools for the detection of endoglucanases as described in (section 2.6.2). Pools F8 and F12 in plate 2 of the fosmid library showed slightly higher absorbance compared to the negative control (Table 5.3).

5.4.3 AZCL-HE-Cellulose assay

Broth-based AZCL-HE-Cellulose assay was performed on the lake-derived fosmid libraries in 96-well format to detect which pools of clones consisted of individual clones producing endoglucanases, as described in (section 2.6.3). High copy number of the fosmids was induced using AutoInduction solution before performing the assays. Pool A7 in plate 2 was the only standout positive within the lake landfill library, with an absorbance value more than double of the negative control (Table 5.4). No other positives were identified from the lake library.

Table 5.2 Absorbance measured at 590 nm in 96-well format following AZCL-Xylan assay on lake-derived fosmid library. (A) Plate 1. Plate average 0.066. (B) Plate 2. Plate average 0.065.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.045	0.047	0.049	0.044	0.047	0.040	0.047	0.046	0.046	0.048	0.057	0.041
B	0.058	0.051	0.066	0.049	0.052	0.050	0.059	0.052	0.059	0.055	0.047	0.030
C	0.060	0.054	0.079	0.053	0.052	0.055	0.056	0.063	0.057	0.058	0.058	0.054
D	0.059	0.055	0.055	0.074	0.052	0.059	0.062	0.065	0.053	0.058	0.063	0.068
E	0.055	0.056	0.058	0.053	0.060	0.054	0.059	0.061	0.054	0.066	0.065	0.047
F	0.059	0.058	0.056	0.054	0.048	0.055	0.062	0.071	0.058	0.058	0.061	0.046
G	0.060	0.061	0.062	0.058	0.054	0.052	0.057	0.061	0.059	0.052	0.062	0.051
H	0.065	0.065	0.059	0.068	0.137	0.070	0.078	0.073	0.077	0.065	0.060	0.048

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.052	0.059	0.056	0.049	0.058	0.052	0.049	0.053	0.051	0.086	0.056	0.039
B	0.058	0.057	0.064	0.061	0.109	0.054	0.056	0.061	0.055	0.057	0.058	0.062
C	0.073	0.087	0.059	0.093	0.062	0.057	0.057	0.059	0.050	0.062	0.055	0.065
D	0.096	0.064	0.078	0.058	0.057	0.061	0.060	0.054	0.052	0.061	0.053	0.067
E	0.066	0.066	0.061	0.058	0.058	0.055	0.057	0.053	0.052	0.053	0.045	0.066
F	0.074	0.071	0.080	0.064	0.051	0.053	0.054	0.056	0.050	0.053	0.053	0.069
G	0.067	0.063	0.077	0.062	0.068	0.055	0.052	0.055	0.049	0.058	0.058	0.068
H	0.061	0.106	0.081	0.072	0.065	0.065	0.065	0.059	0.066	0.077	0.071	0.057

Table 5.3 Absorbance measured at 410 nm in 96-well format following pNPC assay on lake-derived fosmid library. (A) Plate 1. Plate average 0.083. (B) Plate 2. Plate average 0.077.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.086	0.084	0.083	0.086	0.078	0.081	0.078	0.086	0.077	0.080	0.082	0.085
B	0.086	0.091	0.087	0.084	0.068	0.078	0.070	0.073	0.079	0.069	0.074	0.075
C	0.084	0.084	0.099	0.101	0.075	0.092	0.070	0.078	0.070	0.072	0.067	0.078
D	0.084	0.077	0.097	0.096	0.074	0.079	0.072	0.079	0.072	0.070	0.071	0.098
E	0.111	0.090	0.088	0.091	0.072	0.075	0.067	0.095	0.091	0.073	0.070	0.085
F	0.081	0.081	0.080	0.091	0.075	0.078	0.072	0.100	0.081	0.080	0.069	0.084
G	0.085	0.077	0.076	0.105	0.079	0.076	0.070	0.086	0.081	0.073	0.072	0.084
H	0.082	0.081	0.087	0.091	0.087	0.091	0.088	0.086	0.085	0.085	0.081	0.088

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.077	0.078	0.085	0.080	0.077	0.081	0.083	0.079	0.081	0.115	0.081	0.077
B	0.075	0.104	0.084	0.076	0.074	0.075	0.073	0.081	0.077	0.087	0.084	0.082
C	0.075	0.079	0.075	0.094	0.075	0.090	0.081	0.076	0.074	0.083	0.103	0.085
D	0.083	0.073	0.073	0.075	0.072	0.078	0.091	0.077	0.072	0.080	0.085	0.080
E	0.084	0.075	0.070	0.071	0.074	0.071	0.081	0.070	0.069	0.094	0.083	0.078
F	0.079	0.073	0.070	0.074	0.067	0.073	0.075	0.144	0.071	0.083	0.078	0.148
G	0.075	0.070	0.071	0.092	0.071	0.074	0.071	0.075	0.074	0.136	0.075	0.071
H	0.076	0.068	0.074	0.072	0.069	0.072	0.075	0.073	0.068	0.068	0.062	0.074

Table 5.4 Absorbance measured at 590 nm in 96-well format following AZCL-HE-cellulose assay on lake-derived fosmid library. (A) Plate 1. Plate average 0.052. (B) Plate 2. Plate average 0.047.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.049	0.055	0.047	0.047	0.052	0.052	0.049	0.053	0.048	0.055	0.050	0.056
B	0.051	0.060	0.048	0.050	0.050	0.051	0.054	0.066	0.051	0.055	0.051	0.056
C	0.057	0.058	0.049	0.049	0.049	0.055	0.053	0.054	0.052	0.053	0.049	0.056
D	0.052	0.055	0.046	0.053	0.049	0.054	0.049	0.052	0.048	0.054	0.049	0.054
E	0.051	0.060	0.047	0.048	0.059	0.056	0.049	0.052	0.048	0.058	0.051	0.055
F	0.052	0.058	0.047	0.050	0.053	0.053	0.050	0.052	0.049	0.057	0.051	0.054
G	0.050	0.056	0.072	0.046	0.053	0.053	0.049	0.048	0.048	0.052	0.050	0.052
H	0.051	0.053	0.054	0.064	0.049	0.051	0.048	0.051	0.047	0.051	0.050	0.052

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.058	0.049	0.055	0.047	0.053	0.047	0.105	0.044	0.048	0.055	0.051	0.054
B	0.058	0.044	0.046	0.046	0.049	0.050	0.047	0.047	0.046	0.046	0.049	0.053
C	0.053	0.043	0.045	0.045	0.051	0.051	0.048	0.047	0.045	0.048	0.051	0.056
D	0.057	0.043	0.054	0.047	0.049	0.049	0.048	0.050	0.045	0.048	0.052	0.056
E	0.047	0.050	0.057	0.046	0.054	0.053	0.045	0.053	0.044	0.049	0.053	0.058
F	0.051	0.047	0.053	0.053	0.047	0.051	0.045	0.052	0.041	0.047	0.049	0.065
G	0.047	0.044	0.045	0.051	0.044	0.048	0.043	0.050	0.044	0.053	0.051	0.051
H	0.047	0.045	0.049	0.048	0.042	0.043	0.043	0.050	0.045	0.049	0.059	0.049

5.4.4 Congo red assay

The Congo red assay is usually performed by flooding CMC agar culture plates with the Congo red dye, followed by washing with NaCl solution for the isolation of distinct positive clones. However, the Congo red dye is toxic to the bacterial cells and as such, the traditional method is only useful in pointing out which individual colony is positive for expressed endoglucanases. In order to ensure the viability of the identified positives, Congo red assay was performed as described in (section 2.6.1). Sterile 0.45 μm nitrocellulose membranes (PALL) were placed on the surface of CMC agar plates containing chloramphenicol and AutoInduction solution before cells from an individual pool from the fosmid libraries were plated onto the membrane. The pores of the membrane allow for transfer of nutrients to the cells, as well as the transfer of any expressed enzyme onto the media, albeit at a slightly slower rate than usual. After ~ 42 hours incubation, the membranes containing the bacterial colonies were removed before the assay was performed on the plates. The assay, performed on duplicate plates with ~ 350-400 clones on each plate, was used to screen all pools of clones from the lake fosmid library. A positive each from pools A7 and E2 in plate 2 of the library displayed zones of clearance, and was subsequently isolated into pure culture.

5.4.5 Endoglucanase positives from lake-derived fosmid library

The data from the AZCL-HE-Cellulose assay on pools pointed towards the presence of a positive clone within pool A7 in plate 2, which was isolated using the Congo red assay (henceforth called P2A7). A further positive clone was isolated using the Congo red assay from pool E2 in plate 2 (henceforth called P2E2).

5.4.5.1 Extraction of fosmid DNA and sequencing of endoglucanase positives

Both the isolated clones were propagated in fresh freezing medium before fosmid DNA was extracted from P2A7 and P2E2 using the FosmidMAX DNA Purification kit (Epicentre) following the manufacturer's protocol. Roughly 200 ng of fosmid DNA from each clone was processed using the Nextera XT DNA Library Preparation Kit (Illumina) by the Centre for Genomic Research (CGR) to generate barcoded samples for a 2x250 bp paired-end run on the Illumina MiSeq.

5.4.5.2 Fosmid assembly, ORF prediction and functional annotation

The web-based bioinformatics platform Galaxy (www.usegalaxy.org) was used to join the pair-ended reads from each sample using the 'FASTQ joiner' tool (Blankenberg *et al*., 2010), and the sequences were converted into fasta format using the 'FASTQ to FASTA' converter tool. A total of 753,072 and 550,909 sequences from P2A7 and P2E2 fosmids, respectively, were uploaded onto Geneious version 7.1 (www.geneious.com, Kearse *et al*., 2012) for the assembly of the reads to generate a consensus sequence for the two fosmids. While P2E2 was assembled into two main contigs of length ~ 20 kb and ~ 10 kb, attempts to assemble P2A7 yielded many large contigs ranging from length ~ 80 kb to < 1 kb. The 2 contigs from P2E2 and the 50 longest contigs (length ~ 80-6.5 kb) from P2A7 were uploaded on to MetaGeneMark for ORF prediction (Zhu *et al*., 2010). All nucleotide sequences from the predicted ORFs were then searched against the Pfam database (Finn *et al*., 2014) to determine the identity of the proteins encoded on the two fosmids.

Two ORFs from the P2E2 fosmid were determined to contain GH domains (DNA sequences presented in Appendix A). Gene P2E21 was 1,122 nucleotides long,

demonstrated 42% identity to a GH10 domain, and was predicted to have endoxylanase activity. Gene P2E22 on the other hand was 1,221 nucleotides long, demonstrated 22% identity to a GH5 domain, and was predicted to function as an endoglucanase. It was unexpected that P2E2 from lake library did not demonstrate a positive reaction when the broth-based AZCL-Xylan assay was performed on the lake fosmid library, as sequencing result shows the presence of an endoxylanase gene. It is however, possible that the gene might not be transcriptionally active due to issues with codon usage or promoter recognition in a heterologous host, or that it is under the regulation of a weak promoter. It is also possible that the gene product does not undergo the necessary post-translational modifications for the enzyme to be active.

A BlastX search against the NCBI nr database (Altschul *et al*., 1990), with an E value cut-off of 0.001, suggested that the sequences most likely belonged to an environmental strain of *Treponema*. Members of this genus are obligate or facultative anaerobic bacteria belonging to class Spirochetes, and have been reported to be free-living in a variety of aquatic habitats such as sediments and water column of lakes and ponds (Veldkamp, 1960), or as host-associated commensals in the intestinal tracts of ruminants as well as termites (Bryant, 1952; Ohkuma & Kudo, 1996). They are able to ferment pectin, inulin, sucrose, L-arabinose as well as cellobiose to generate acetic and formic acids (Wojciechowicz & Ziolecki, 1979).

Although Treponemes have not been reported to play a specific role in lignocellulose degradation, they have been found attached to the fibrous plant feed in the rumen, where their interaction with cellulolytic bacteria such as *Bacteroides succinogenes*, *Ruminococcus albus, Clostridium thermocellum* and *Fibrobacter succinogenes* has been documented to aid cellulose degradation in that environment (Stanton & Canale-Parola, 1980; Kudo *et al*., 1987). It has been suggested that this is

due to utilisation of cellobiose by the spirochete, relieving the feedback inhibition of cellulase activity by the cellulose degradation product (Leschine, 1995). However, Bekele *et al*. (2011) reported that a large proportion of Treponemes remain uncultured, and that these spirochetes appear to have distinct members that play a role in the active degradation of insoluble plant matter such as hay. The isolation of an endoglucanase and an endoxylanase most likely belonging to an environmental strain of *Treponema* from a lake-derived fosmid library seems to support such a report.

No glycoside hydrolases were found to be present in the 50 P2A7 contigs searched against the Pfam database, and the BlastX results surprisingly pointed towards the ORFs having closest hits to *E. coli* sequences. It would appear that the fosmid DNA extracted from the P2A7 isolate was contaminated with some host *E. coli* DNA. This would explain not only the presence of vast amounts of *E. coli* gene sequences in the dataset, but also the peculiar assembly of the fosmid that was observed, as some contigs generated were longer than ~ 40-45 kb.

One way of solving the issue is to perform a BlastX against a GH database such as CAZy (Canatrel *et al*., 2009) using the P2A7 fosmid sequences as the query. Any GH hits obtained can then be used to perform a BlastX search against the NCBI nr database to determine the phylogenetic origin of the sequences. Creating an *E. coli* sequence database to perform a Blast search of the P2A7 fosmid sequences against it can potentially solve the issue too, as any sequences without a hit would have been encoded on the fosmid and could subsequently be separated out. ORFs could then be predicted in those sequences, followed by a search against the Pfam database to determine the presence of any GH domains. Alternatively, fosmid DNA could be re-extracted from the P2A7 isolate and sequenced again.

5.5 Functional screening of the landfill-derived fosmid library for isolation of lignocellulosic enzymes

Following the successful screening of the lake-derived fosmid library, pNPC assay, AZCL-HE-Cellulose assay, AZCL-Xylan assay and Congo red assay were also used to screen the landfill fosmid library. All assays were performed using commercially available cellulase or xylanase (Sigma-Aldrich) as positive control, and a culture of *E. coli* DH5α as negative control.

5.5.1 AZCL-Xylan assay

Broth-based AZCL-Xylan assay was performed on pools of clones to detect expression of endoxylanase as described before. Only pools B2 and C9 from plate 2 were determined to be positive from the landfill fosmid library, whereas no positives were observed from plate 1 (Table 5.5).

5.5.2 pNPC assay

Detection of endoglucanases in pools of clones was attempted using the pNPC assay, as described before. No pools from either plate demonstrated any significantly higher absorbance in the landfill-derived library (Table 5.6).

5.5.3 AZCL-HE-Cellulose assay

Endoglucanase expression in pools of induced fosmid clones was also tested using the broth-based AZCL-HE-Cellulose assay. By comparison to the lake fosmid library, the landfill fosmid library contained many positives. Pools A2, F4, G6, G11, H3, H4, H5, H6, H8, H10, H11 and H12 in plate 1 and pools C12, F4, F12 and G1 in

plate 2 were strongly positive, while pools E4, E5, E11, F11, G4, G9 and H2 in plate 1 displayed slightly higher absorbance compared to the negative control (Table 5.7).

Table 5.5 Absorbance measured at 590 nm in 96-well format following AZCL-Xylan assay on landfill-derived fosmid library. (A) Plate 1. Plate average 0.054. (B) Plate 2. Plate average 0.065. The last 16 wells of the plate (yellow) do not contain any clones.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.039	0.044	0.040	0.043	0.056	0.070	0.073	0.060	0.045	0.046	0.039	0.036
B	0.048	0.050	0.059	0.088	0.082	0.082	0.074	0.070	0.057	0.049	0.048	0.045
C	0.047	0.049	0.056	0.092	0.074	0.079	0.064	0.070	0.055	0.055	0.049	0.044
D	0.050	0.049	0.067	0.080	0.073	0.067	0.050	0.046	0.050	0.051	0.045	0.038
E	0.048	0.047	0.065	0.083	0.071	0.049	0.046	0.047	0.047	0.053	0.052	0.041
F	0.059	0.055	0.068	0.073	0.051	0.057	0.047	0.044	0.043	0.052	0.053	0.042
G	0.049	0.045	0.046	0.048	0.048	0.053	0.045	0.044	0.042	0.057	0.045	0.037
H	0.065	0.050	0.044	0.044	0.057	0.049	0.038	0.045	0.044	0.046	0.039	0.041

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.057	0.057	0.069	0.050	0.051	0.058	0.053	0.073	0.072	0.047	0.053	0.050
B	0.081	0.100	0.055	0.058	0.061	0.063	0.066	0.076	0.090	0.080	0.085	0.050
C	0.055	0.051	0.064	0.056	0.059	0.061	0.058	0.072	0.134	0.084	0.068	0.055
D	0.058	0.055	0.060	0.051	0.059	0.057	0.059	0.065	0.060	0.069	0.064	0.057
E	0.055	0.051	0.058	0.057	0.061	0.061	0.073	0.079	0.082	0.077	0.062	0.051
F	0.058	0.054	0.054	0.059	0.067	0.065	0.078	0.081	0.067	0.059	0.052	0.049
G	0.063	0.063	0.080	0.072	0.077	0.055	0.094	0.087	0.061	0.052	0.046	0.052
H	0.046	0.051	0.062	0.060	0.052	0.082	0.069	0.064	0.055	0.061	0.049	0.050

Table 5.6 Absorbance measured at 410 nm in 96-well format following pNPC assay on landfill-derived fosmid library. (A) Plate 1. Plate average 0.170. (B) Plate 2. Plate average 0.175. The last 16 wells of the plate (yellow) do not contain any clones.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.178	0.170	0.163	0.173	0.171	0.170	0.176	0.171	0.166	0.165	0.158	0.148
B	0.174	0.172	0.166	0.167	0.166	0.161	0.165	0.165	0.175	0.162	0.127	0.165
C	0.183	0.182	0.172	0.175	0.166	0.161	0.159	0.162	0.163	0.164	0.145	0.163
D	0.184	0.167	0.170	0.174	0.165	0.160	0.166	0.170	0.180	0.169	0.146	0.140
E	0.185	0.178	0.165	0.174	0.170	0.161	0.161	0.161	0.173	0.148	0.173	0.160
F	0.179	0.167	0.164	0.170	0.169	0.167	0.165	0.165	0.125	0.161	0.166	0.171
G	0.180	0.180	0.164	0.167	0.174	0.169	0.165	0.168	0.171	0.135	0.160	0.186
H	0.187	0.187	0.176	0.173	0.176	0.175	0.180	0.159	0.158	0.182	0.168	0.198

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.225	0.212	0.210	0.199	0.200	0.206	0.196	0.206	0.202	0.198	0.207	0.206
B	0.209	0.176	0.175	0.175	0.175	0.180	0.174	0.180	0.178	0.171	0.172	0.169
C	0.165	0.171	0.170	0.164	0.177	0.175	0.172	0.164	0.169	0.160	0.161	0.170
D	0.191	0.170	0.169	0.170	0.173	0.170	0.157	0.167	0.166	0.164	0.170	0.176
E	0.175	0.173	0.162	0.171	0.170	0.170	0.158	0.167	0.145	0.169	0.168	0.175
F	0.195	0.175	0.173	0.163	0.169	0.170	0.166	0.171	0.158	0.174	0.173	0.151
G	0.170	0.175	0.166	0.173	0.172	0.163	0.174	0.170	0.172	0.174	0.172	0.179
H	0.179	0.168	0.172	0.169	0.168	0.168	0.172	0.174	0.178	0.181	0.173	0.174

Table 5.7 Absorbance measured at 590 nm in 96-well format following AZCL-HE-Cellulose assay on landfill-derived fosmid library. (A) Plate 1. Plate average 0.064. (B) Plate 2. Plate average 0.067. The last 16 wells of the plate (yellow) do not contain any clones.

A.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.056	0.111	0.079	0.079	0.076	0.084	0.043	0.063	0.056	0.037	0.052	0.065
B	0.044	0.030	0.066	0.078	0.058	0.051	0.064	0.058	0.066	0.086	0.081	0.057
C	0.047	0.035	0.079	0.080	0.068	0.081	0.078	0.081	0.068	0.075	0.060	0.057
D	0.052	0.041	0.073	0.078	0.071	0.077	0.073	0.086	0.070	0.072	0.083	0.060
E	0.048	0.042	0.081	0.093	0.066	0.095	0.083	0.069	0.082	0.083	0.092	0.067
F	0.050	0.049	0.068	0.100	0.081	0.078	0.078	0.074	0.074	0.075	0.099	0.064
G	0.059	0.049	0.068	0.095	0.070	0.105	0.085	0.077	0.098	0.088	0.101	0.057
H	0.082	0.095	0.100	0.120	0.109	0.111	0.069	0.117	0.079	0.140	0.125	0.110

B.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.078	0.061	0.073	0.088	0.052	0.080	0.056	0.046	0.064	0.044	0.079	0.068
B	0.080	0.064	0.083	0.081	0.062	0.066	0.063	0.063	0.086	0.083	0.081	0.076
C	0.065	0.067	0.067	0.067	0.060	0.060	0.065	0.085	0.061	0.061	0.079	0.108
D	0.068	0.065	0.060	0.078	0.071	0.070	0.069	0.075	0.062	0.070	0.077	0.087
E	0.084	0.081	0.070	0.088	0.077	0.067	0.073	0.065	0.070	0.058	0.088	0.082
F	0.060	0.070	0.076	0.114	0.073	0.067	0.080	0.071	0.085	0.086	0.076	0.127
G	0.116	0.070	0.066	0.085	0.072	0.084	0.070	0.081	0.039	0.039	0.045	0.047
H	0.044	0.042	0.043	0.044	0.047	0.044	0.045	0.039	0.045	0.048	0.044	0.041

5.5.4 Congo red assay

The membrane-based Congo-red assay was performed on duplicate plates for the isolation of individual clones expressing endoglucanases as described before. However, limitations imposed by time meant only the pools found to be strongly positive using the AZCL-HE-Cellulose assay, i.e. pools A2, F4, G6, G11, H3, H4, H5, H6, H8, H10, H11 and H12 in plate 1 and pools C12, F4, F12 and G1 in plate 2, were screened from the landfill library. No positives were detected from the few pools screened. This was surprising as the only pool producing a positive AZCL-HE-Cellulose assay from the lake fosmid library did indeed yield a positive isolate using the Congo red assay.

5.5.5 Troubleshooting and future work

In order to troubleshoot the inability to isolate clones expressing endoglucanases using Congo red method, the assay was repeated in quadruplicate on the aforementioned pools to ensure that enough clones were screened to provide sufficient coverage. Once again, no positives were detected. The broth-based AZCL-HE-Cellulose assay was also repeated to check if the apparent positive pools could be replicated. Curiously, the results were vastly different as only 3 pools were found to produce a blue reaction in both the replicates, as many different pools produced a positive reaction when replicated. The reasons for this are unknown. However, it is unlike that it is a result of substrate disintegration, as the negative controls were not found to display significantly higher absorbance readings compared to previously conducted assays. One of the pools found to be positive in both replicates of the assay, P2F12, was chosen to produce a library of individual clones for the detection of the discrete positive clone within the pool. The inoculum from the pool was spread plated

onto LB agar plates supplemented with 12.5 μg/ml chloramphenicol to yield ~ 800 clones. Each separate clone was picked and propagated into individual wells of eight 1 ml MasterBlock 96-well plates (Greiner) to yield a constitutive library of 768 clones from pool P2F12. Since no positive isolates could be identified using the Congo red assay, broth-based AZCL-HE-Cellulose assay was performed on the 768 clones in 96-well format following induction of high copy number of fosmid. 21 clones were found to produce a positive reaction to the assay. In order to confirm endoglucanase expression, the 21 isolates were screened using the Congo red method using membranes, as described before. Disappointingly, no zones of clearance were observed on the CMC agar plates from any of the isolates, and therefore no endoglucanase expression was confirmed from the pool P2F12. The remaining 2 positive pools could not be screened following this method due to lack of time.

The pool P2F12 was also screened using the agar plate-based AZCL-HE-Cellulose assay (Nguyen *et al*., 2012) as described in (section 2.6.4). The assay was performed in quadruplicate and the spread plates containing ~ 300 clones each were overlaid with top agar (0.7% w/v) containing the dye cross-linked substrate, before being incubated overnight at 37ºC. No clones were determined to produce a positive reaction, as no blue dye was released near any of the clones. Since broth-based AZCL-HE-Cellulose assay provided positives, this result was unexpected. In order to troubleshoot, agar plate-based assay was also performed on the endoglucanase positive P2A7 isolated from the lake-derived fosmid library. Interestingly, no clones were determined to produce a positive reaction. Time constraints mean that all pools that appeared to be produce a positive reaction to the broth-based AZCL-HE-Cellulose assay from the landfill library could not be screened using the agar-based method. A possible explanation for failure to observe any positives is that the colonies and the

substrate present in the top agar are not in close enough proximity for any expressed enzyme to hydrolyse the substrate.

Nguyen *et al*. (2012) used ENZhance permeability kit (BIOTEC) while performing the agar plate-based AZCL-HE-Cellulose assay in order to increase the permeability of the bacterial clone cell membranes to facilitate enzyme 'leakage', and subsequently boost the probability of finding expressed enzymes of interest. The method can be used to screen both the lake and the landfill fosmid libraries in their entirety, as the ENZhance reagent does not lyse bacterial cells completely and hence any positives detected can be propagated for further analysis.

Cell lysis can also be incorporated into the pNPC assay, as performed by Mewis *et al.* (2011). A lysis mix consisting of 10% Triton X-100, 100 mM Tris and 10 mM EDTA was added to the assay mix and led to the identification of a clone positive for cellulase expression. The same cocktail of lysing agents can theoretically be used in the broth-based AZCL-HE-Cellulose and AZCL-Xylan assays too. A crude method to potentially induce cell lysis during broth-based assays could be to increase the assay incubation time from overnight to up to a week in order to induce cell death, before performing spectrophotometry.

Measuring relative absorbance by measuring difference in absorbance over the plate average when performing broth-based assays is considered to be a more robust method for the identification of cellulase positive clones (Mewis *et al*., 2011). This is clearly demonstrated by the pNPC assay data where absorbance values obtained from the landfill fosmid library (Table 5.6) are higher than those from the lake fosmid library (Table 5.3). Yet, no clear positives were observed from the landfill library, as this higher absorbance is likely to be an artefact of inherent plate variation between each separate assay run performed at a different time.

Both the landfill and the lake fosmid libraries should be fully screened for the expression of xylanases using agar-based methods in order to isolate individual positives. The method for this is relatively straightforward, as LB agar plates supplemented with 0.2% (w/v) xylan from oat spelt appear to have a slightly cloudy appearance (Prade, 1996). Clones expressing xylanases produce a clear zone around them for easy identification. Alternatively, Congo red assay can also be used for the screening of endoxylanases, where 0.2% (w/v) xylan from beechwood is added to LB agar. The xylan reacts with the Congo red dye to provide a red colour to the plate, and any clones expressing endoxylanases can be detected visually by virtue of having a zone of clearance around them.

Congo red screening of every pool from the landfill fosmid library should be carried out. It is entirely possible that clones expressing endoglucanases can be isolated, since Congo red screening of the lake library led to the isolation of a positive clone from a pool that was not identified to be positive using the AZCL-HE-Cellulose screen. Furthermore, given the inconsistent results obtained from the AZCL-HE-Cellulose assay, Congo red screening can be regarded as the primary fosmid library screening method for the mining of endoglucanases not only because of the fact that it can pin-point discrete clones, but also due to the number of times it has yielded positives in other studies (Voget *et al*., 2006; Zhang *et al*., 2011; Duan *et al*., 2009; Kim *et al*., 2011).

5.6 Discussion

The HMW DNA extraction method modified from Neufeld *et al*. (2007) yielded the majority of the DNA used for the production of the landfill fosmid library, whilst also providing DNA that was cleaner than that obtained from the Mete-G-Nome DNA

Extraction Kit (Epicentre). Nonetheless, DNA extracted from both the methods was used and this might prove to be beneficial in achieving improved coverage of the microbial community metagenome, as it has been shown recently that diversification of extraction methods can lead to up to an 80% increase in the genetic diversity recovered (Delmont *et al*., 2011).

The number of clones that constitute the lake and the landfill fosmid libraries are consistent with the literature, especially where sampling has been done from complex environments consisting of highly heterogeneous microbial communities (Kakirde *et al*., 2010). It was, however, surprising that despite the implementation of sample enrichment, the hit rate for the detection of endoglucanases was very low in the lake fosmid library. A full and comprehensive screening of the landfill fosmid library would be required before conclusions can be drawn regarding the success of the enrichment method.

The process of shotgun metagenomic cloning is random and due to environmental microbial communities consisting of an abundant and highly diverse population, it is not possible to guarantee extensive coverage of the metagenome representing such a population. This means that it is possible some significant components of the metagenome, including specific genes of interest, may not be present in the clone library. Moreover, the extracted DNA pool can be further biased by the effect of sampling, cell separation and the lysis method employed (Ekkers *et al*., 2012). To that effect, the use of more than one DNA extraction method employed here might prove to be beneficial in achieving improved coverage of the microbial metagenome as it has been shown that diversification of extraction methods can lead to up to an 80% increase in the genetic diversity recovered (Delmont *et al*., 2011).

Despite the promise of genetic novelty, function-based metagenomic studies often do not succeed in returning products of sufficient novelty for application in biotechnological processes (Singh & Macdonald, 2010). This can be attributed to the low level of gene expression often observed in a heterologous host, while function-based screening is also prone to the rediscovery of already known functions (Binga *et al*., 2008). A number of factors regulate whether a foreign gene product can be expressed in a heterologous host, and the degree of non-expression varies between different gene-host combinations, making it very difficult to predict the percentage of genes from an environmental fosmid library being expressed in a given host. One of the major bottlenecks is imposed by codon usage bias, whereby most microorganisms have a certain preference for the usage of specific codons when encoding signals for initiating and terminating protein translation (Kudla *et al*., 2009). Both Goodarzi *et al*. (2008) and Kudla *et al*. (2009) demonstrated that the preference for start codons displays interspecific variation, leading to difference in expression levels that can vary by 250-fold. Since very little is known regarding the community and gene composition in most shotgun metagenomic clone library studies, appropriate host expression machinery is unlikely to be chosen and optimised for protein production. This is significant when functional screening of metagenomic libraries is performed, as the usage of optimal codons influences translation (Sorensen *et al*., 1989), protein folding (Zalucki *et al*., 2009) and secretion (Zalucki & Jennings, 2007).

It is also worth pointing out that non-catalytic proteins such as carbohydrate-binding domains, which are essential components of cellulosomes, are unable to be detected using such expression screens. Since the objective of the study was to mine for novel glycoside hydrolases from any possible microbial source, enhancements for decreasing library complexity such as biasing genetic material based on the G-C content

(Holben, 2011) or pre-selecting target microbes to suit the host expression machinery were not practical considerations. Gabor *et al*. (2004) reported that only ~40% of gene products can be detected by expression screening when random cloning of DNA from 32 prokaryotic genomes was performed in an *E. coli* host. However, this number is expected to be much lower as this study only relied on theoretical bioinformatics considerations, and did not take into account the multitude of factors that are required for absolute gene expression, including the presence of co-factors that are involved in protein folding and secretion.

Other factors that play a potentially crucial role in the lack of enzyme expression and detection include lack of initiation factors leading to erroneous promoter recognition and ribosomal entry, lack of post-translational modifications causing improper protein folding, enzymatic breakdown of the foreign gene product and inclusion body formation, toxicity of the gene product, as well as the inability of the host to secrete the expressed protein (Ekkers *et al*., 2012). The issue is further exaggerated when screening for enzymes encoded by eukaryotes in a prokaryotic expression system. Therefore, it is quite likely that a significantly higher number of genes encoding glycoside hydrolases were cloned into the two fosmid libraries than the assay results suggest. The metatranscriptomic data generated from parallel samples is thus an exceptional resource as sequences encoding glycoside hydrolases have been identified. These can be employed for the generation of PCR primers and probes to allow for sequence-based screening of the fosmid libraries to overcome some of the aforementioned issues with functional screening, particularly those related to post-translational modifications and secretion.

Engineering the host transcription and translation machinery to increase host recognition of ribosome binding sites based on the expected prevalence of genes in a

microbial community (Bernstein *et al.*, 2007) and enhancing protein folding by increasing the co-expression of chaperone proteins (Ferrer *et al*., 2004) have been suggested as potential ways by which expression of foreign proteins can be boosted in a heterologous host. Bayer *et al*. (2009) used sequence information to codon optimise genes potentially related to methyl halide transferases for expression in *E. coli* and reported the successful expression and methyl halide transferase activity from 83 out of the 89 genes. However, these methods remain technically challenging, time consuming, as well as expensive, particularly due to the requirement for synthesis of all the target genes.

E. coli has been used predominantly as the cloning host for expression screening of metagenomic libraries, as an extensive tool kit is available for the genetic manipulation of this well characterised model microorganism. The relatively narrow-range expression profile observed in *E. coli* is a major factor leading to the functional metagenomics screening bottlenecks encountered. As such, the use of multiple hosts for expression is becoming increasingly common, as it confers the advantage of an increased chance of protein production, whilst bypassing issues with accelerated enzymatic breakdown and toxicity of the gene product within an individual host (Ekkers *et al*., 2012). Shuttle vectors with broad host range have been used in the past for this purpose, and integrating features that include inducible copy number and the ability to incorporate HMW DNA benefits their design. Aakvik *et al*. (2009) used a BAC vector that could be successfully transferred into *Pseudomonas fluorescens* and *Xanthomonas campestris*, whereas Craig *et al*. (2010) constructed metagenomic libraries in a broad host-range cosmid that allowed screening to be performed in *Agrobacterium tumefaciens*, *Burkholderia graminis*, *Caulobacter vibrioides*, *E. coli*, *Pseudomonas putida* and *Ralstonia metallidurans*. The latter study is significant as

little overlap was reported between the highly diverse expression profiles observed in the different hosts, strengthening claims that the choice of expression host can yield significantly different expression data from the same metagenomic library.

Direct detection of gene expression products using assays incorporating substrates such as Congo red dye, AZCL-HE cellulose and AZCL xylan is relatively 'low-tech' and allows for certain aspects of the screening to be coupled with high throughput technologies. However, direct detection screening of large clone libraries consisting of 50,000-100,000 clones is labour intensive and can be disadvantageous given its low resolution due to the inability to identify positive clones displaying low level of expression. Using colony picking robots and droplet-based microfluidic approaches in conjunction with 384-well plates and microplate readers enhances the reliability and comparability of assays performed on different libraries, whilst shortening the processing time considerably (Taupp *et al*., 2011; Mewis *et al*., 2011). However, automated technology is not available to most research groups, usually due to the prohibitively high cost of the initial set up.

Another fundamental issue with expression-based screening is that the cloned enzymes will only be expressed intracellularly in the heterologous host due to the lack of signal sequences required for the secretion of the target gene product (Ekkers *et al*., 2012). This is essential for detection by assaying, as the cell membrane might not necessarily be permeable and insufficient 'leakage' of the enzyme might result in false negatives being observed. Forced lysis of the host cells while performing the assay is a possible solution (Bao *et al*., 2011). However, lysing agents must be chosen carefully to not interfere with the substrate-enzyme interaction and also to ensure that the protein product of interest is not denatured by its action. Mechanical cell disruption by processes like sonication is usually labour intensive and time consuming. Li *et al*.

(2007) reported the development and use of a UV-inducible autolytic vector for high throughput screening of expressed β-galactosidase activity. Lysis rate of ~60% or more was observed at 30ºC, yet the rate at 37ºC was much less consistent. This suggests that further research is necessary for the design of a more robust autolytic vector for the purpose of aiding functional screening of metagenomic libraries. The use of reporter genes for high throughput screening has been suggested as it has a number of advantages as the detection of positive clones does not require successful expression of a protein downstream of transcription, and faint expression of the cloned gene still produces enough signal for detection (Schipper *et al*., 2009).

In conclusion, the inconsistency observed in the AZCL-HE-Cellulose assay suggests that all broth-based assays must be replicated in order to determine if the data are consistent. The predominant reason for failure to do so was the lack of time. However, it is also worth pointing out that synthetic substrates used for these assays are expensive. An extensive screening of the two fosmid libraries for glycoside hydrolases is expected to be a time consuming process, especially when only one person is involved in the screening process. Nevertheless, the fosmid libraries represent an excellent resource for the mining of not only glycoside hydrolases, but also other enzymes of interest, given the DNA used for the production of the libraries was extracted from understudied anaerobic environments.

Chapter 6

Development of a suitable method for extracting RNA from microbial communities in hyperalkaline environments

6.1 Background

Although high pH environments such as soda lakes occur naturally on earth, they can also be produced as a result of human activity, particularly due to the generation of waste residues from industries such as lime production, borax and cementitious construction (Burke *et al*., 2012). Hydroxide waste contamination of the surface environment from a lime kiln waste site at Brookbottom, Harpur Hill, Derbyshire, UK has led to the formation of an anthropogenic high pH (>11.0) environment, where *in-situ* pH values of up to pH 12.0 and pH 13.5 have been recorded in the soil profile and the alkaline leachate, respectively (Rout *et al*., 2015).

Rainwater percolation through the CaO containing wastes has led to the generation and subsequent deposition of an alkaline leachate (pH 12.0-13.0), containing $Ca(OH)_2$, in the valley downstream. Contact with atmospheric CO_2 has led to the formation of a $CaCO_3$-rich calcite precipitate deposit (tufa), which has in-filled the valley floor and is encroaching upon the adjacent farmland (Figure 6.1). Previous research has reported the presence of anaerobic, alkaliphilic microbial communities within the organic substrate-rich layer of soil found at the site (Burke *et al*., 2012). As such, the identification and isolation of lignocellulose-degrading enzymes capable of withstanding and operating in highly alkaline conditions has significance in industrial applications such as consolidated biomass processing, particularly as alkaline peroxide

(Gould, 1985) and ammonia fibre explosion (Wyman *et al*., 2009) are commonly utilised lignocellulose pre-treatment methods in industry.

Understanding the microbiology of alkaline cellulose degradation has further significance in the nuclear industry, as highly alkaline and anoxic conditions are also expected to be typical of radioactive waste repositories, where a cement-based disposal concept is currently under review by the Nuclear Decommissioning Authority (NDA) in the UK. Briefly, the current proposal lists long-lived intermediate- and low-level wastes (ILW/LLW) to be disposed of in vault environments backfilled with cementitious materials (Nirex Reference Vault Backfill, NRVB) after their chemical conditioning (Nirex report, 2003; Nirex report, 2005). Such conditioning, supplemented by groundwater ingress and corrosion processes, ensures the prevalence of anoxic, alkaline conditions (pH 12.5-13.5) rich in Ca^{2+} ions in the repository for long periods post closure. Organic materials present in the waste, including paper, wood and cotton, are susceptible to alkaline degradation and this phenomenon is of significant interest as cellulose degradation products (CDP) are able to form complexes with disposed radioelements, facilitating their migration and potentially enhancing the associated radiological risks (Greenfield *et al*., 1997; Heath & Williams, 2005).

Figure 6.1 Calcite-rich tufa deposits at Brookbottom valley floor (arrowed). Decomposing plant matter can be seen engulfed by the alkaline leachate.

Alkaline cellulose degradation typically involves three main phases that control the rate and the overall extent of the decomposition (Glaus & Van Loon, 2008). The first phase consists of the rapid 'peeling' of monomeric end groups from the reducing end, along with hydroxide catalysed 'mid-chain scission' of glycosidic bonds that generates new reducing ends randomly along the cellulose chain. This is followed by a slower second phase where the rate of degradation is controlled by release of new end groups from amorphous segments of cellulose chains as well as the physical access to available end groups. The ceasing of 'mid-chain scission' reactions due to a lack of amorphous cellulose leads to the advent of the final phase, where decomposition is physically inhibited by the high crystallinity of the cellulose fibres. At this point, degradation might completely cease, or it might proceed extremely slowly over the long term due to the transformation of crystalline regions into amorphous regions via currently undefined reaction pathways (Humphreys *et al*., 2010).

CDPs predominantly consist of the soluble organic compounds α- and β-isosaccharinic acid (ISA), whilst short chain fatty acids are also generated in small quantities (Whistler & Bemiller, 1958; Knill & Kennedy, 2003). As such, the rate of chemical degradation of cellulose has been measured almost exclusively as the rate of conversion of insoluble cellulose into alkali soluble organic carbon. Exposure to alpha and gamma radiation enhances alkaline cellulose degradation by increasing 'mid-chain scission' reactions and subsequently reducing the crystallinity of the substrate (Heath & Williams, 2005). Chemical degradation of ISA has been demonstrated to occur only in the presence of oxygen or a similar oxidant (Glaus *et al*., 2008). Despite possessing a similar cellulose composition, environmental conditions within radioactive waste disposal sites are not expected to resemble those associated with landfill sites, mainly

due to factors such as the higher pH and lower water content typically found in the former (Grant *et al*., 1997).

In summary, cellulosic biomass originating from decomposing plant matter constitutes the major source of biodegradable carbon in hyperalkaline lagoons such as Harpur Hill. Chemical degradation of cellulose into ISA at high pH and the ability of the degradation product to mobilise radionuclides (Heath & Williams, 2005), as well as biological degradation of cellulose mediated by industrially-applicable cellulases provide context to this study. For metagenomic studies in such environments, DNA stability is probably not an issue but RNA is extremely susceptible to hydrolysis at high pH (> pH 8.0) due to the presence of 2'-OH group, which contributes towards the cleavage of the RNA backbone by intramolecular hydrolysis of the phosphodiester bond. Given that practically no specialist literature exists outlining a protocol, the aim of the study was to develop a methodology for extracting high quality RNA from mixed microbial communities in hyperalkaline environments. It was hypothesized that development of a suitable RNA extraction method, through neutralization of the highly alkaline sample, would facilitate metatranscriptomic studies into unravelling the structure and function of similar alkaliphilic communities, and aid gene mining for the discovery of novel enzymes. DNA extractions for metagenomic analysis were also performed in this study.

6.2 Sampling and experimental design

The following samples were initially included:

- **Biofilm from incubated cellulose cotton**: Cotton was chemically treated and incubated in the anaerobic segment of the hyperalkaline site by collaborators at the University of Huddersfield, as described in Charles *et al*. (2015). Briefly,

the natural waxes and impurities associated with cotton were removed by treatment with NaOH and phosphate ester detergent, followed by bleaching of the fabric and neutralisation under acetic acid. Approximately 5 g of sterile-treated cotton was placed inside a nylon mesh bag and incubated at the bottom of a 0.5 m borehole drilled into an area of the site impacted by the alkaline leachate. The colonised cotton samples were harvested after a period of 3 months.

- **Cotton biofilm-driven microcosm**: The colonised cotton sample (as described above) was used to set up a microcosm experiment by the same team (Charles *et al*., 2015). Briefly, 1 g of colonised cotton was gently washed with sterile PBS to remove transient microbes. The washed cotton was subsequently suspended in 175 ml of 10% CDP and 90% mineral media that was supplemented with 25 ml of synthesised alkaline CDPs every 2 weeks (Rout *et al*., 2014), before being maintained at pH 11.0 (using 4M NaOH) and a temperature of 20ºC. The cotton was removed once the microcosm reached a total volume of 250 ml. This was followed by a feed and waste cycle being operated in the stirred reaction vessel (British standards institute, 2005) where 25 ml of microcosm contents were replaced by 25 ml of CDP, while the system was maintained anoxically under nitrogen for a period of over 50 weeks before any samples were taken. The continually generated microcosm run-off was the biological sample used for molecular biological analysis.
- **Contaminated soil-driven microcosm**: A microcosm experiment was set up by the research group at University of Huddersfield as described in Rout *et al*. (2015). Briefly, 50 g of soil impacted by the alkaline leachate from the Harpur Hill site was suspended in 250 ml of anaerobic mineral media that was

supplemented with 50 ml of synthesised alkaline CDPs every 2 weeks, before being maintained at pH 11.0 and a temperature of 25ºC. A feed and waste cycle was operated in the stirred reaction vessel as described above once a final volume of 500 ml was achieved, where 50 ml of microcosm contents were replaced by 50 ml of CDP, while the system was maintained anaerobically under nitrogen for a period of over 20 weeks before any samples were taken. As above, the continually generated microcosm run-off was the biological sample used for molecular biological analysis.

Both the microcosm samples were of interest to collaborators at the University of Huddersfield for understanding the mechanism of ISA degradation and assimilation, while the colonised cotton sample was of interest to me as a resource for the discovery of novel cellulases using metatranscriptomics.

6.3 DNA extraction

A vast amount of cementitious matter was found associated with the microcosm samples and visual observation of the incubated cotton sample demonstrated sparse colonisation. As such, a method modified from the Meta-G-Nome DNA Isolation Kit (Epicentre) was chosen for DNA extractions (section 2.2.3). The method relies on displacing cells attached to surfaces using high speed vortex mixing in 1 x PBS (pH 7.0) supplemented with 0.1% Tween 20, followed by differential centrifugation to yield a cell suspension free from particulate matter. The cells were then subjected to DNA extraction according to Griffiths *et al.* (2000) (section 2.2.2). 10 ml of soil-driven and 20 ml of biofilm-driven microcosm run-off was used as starting material for DNA extractions, while the amount of colonised cotton used was 2 g. The DNA yields obtained following Riboshredder RNase treatment and clean-up using Isolate II PCR

and Gel Kit were ~800 ng, ~350 ng and <100 ng from the soil-driven microcosm, biofilm-driven microcosm and the colonised cotton, respectively. By comparison, ~10 μg and ~7.5 μg of DNA could be typically extracted from equivalent amount of landfill leachate and cotton incubated in landfill leachate, respectively (section 3.2.1, section 3.2.2). Agarose gel electrophoresis was performed to determine the quality of the extracted DNA from the microcosms (Figure 6.2A). The typical DNA profile observed demonstrated the DNA to be slightly sheared, with the majority of it present within the size range 600-3000 bp. Despite the lack of a prominent DNA band of ~10 kb, the DNA was deemed to be suitable for Illumina MiSeq sequencing as there was no 'tail' of DNA below the 600 bp mark.

Further extractions were performed and the DNA from 5 g of colonised cotton was pooled together to boost the yield. However, no more than ~250 ng of DNA could be obtained, which was not sufficient to perform conclusive quality control as well as high throughput sequencing. Due to the time constraints involving preparation, incubation and sampling of fresh cotton cellulose baits, a new sample was chosen for the study of cellulases in hyperalkaline environments. Soil rich in decaying lignocellulosic matter was sampled from ~12 inches below the ground from the site at the soil-alkaline leachate interface. The soil was supplemented with pH 13.0 leachate and incubated anaerobically inside a 50 ml centrifuge tube (Greiner) at room temperature in the laboratory (University of Huddersfield) for > 6 months to allow for equilibration and establishment of a stable microbial community. 10 g of the soil-leachate mixture was used for DNA extraction as described above, and a high yield of ~ 7 μg was obtained following RNA removal and the subsequent clean-up. Agarose gel electrophoresis determined the DNA profile to be similar to that obtained from the microcosm samples, with the majority of the DNA present within the size range 600-

3000 bp (Figure 6.2B). Moreover, a small fraction of the DNA could also be observed around the 6-10 kb size range. The extracted DNA was also determined to be suitable for Illumina MiSeq sequencing.

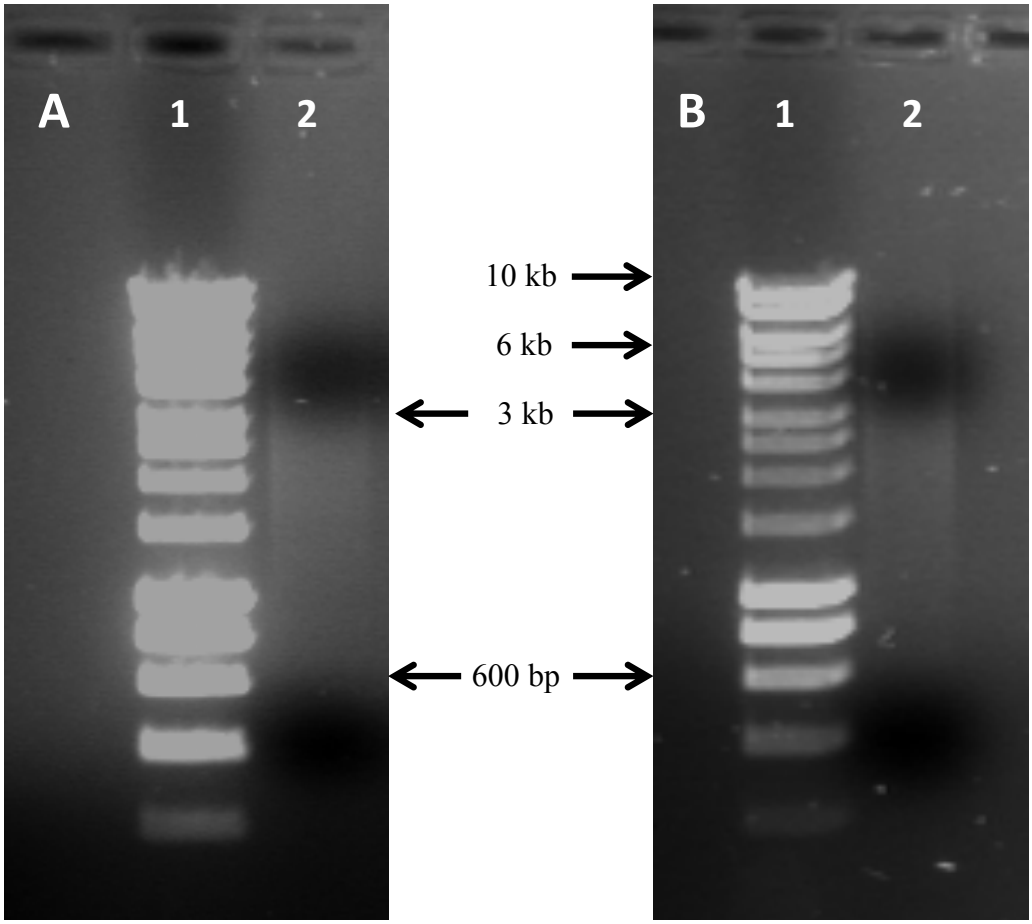


Figure 6.2 Agarose gel electrophoresis of extracted DNA samples. (A) Lane 1, size marker (Hyperladder I, Bioline); lane 2, typical profile of DNA extracted from the soil-driven microcosm runoff. The majority of the DNA can be seen within the size range 600 bp-3 kb. (B) Lane 1, size marker (Hyperladder I, Bioline); lane 2, profile of DNA extracted from the soil-leachate mixture. The majority of the DNA can be seen within the size range 600 bp-3 kb, while a small fraction can also be seen within the size range 6 kb-10 kb.

6.4 RNA extraction

6.4.1 Initial extraction

Extractions were performed on the colonised cotton, soil-driven and biofilm-driven microcosm runoff samples using the same method as used for DNA extractions, albeit following the modifications listed in section 2.2.8 for RNA extractions. Due to the low DNA yields observed previously and the need for relatively high total RNA to be extracted prior to extensive post-processing to obtain high throughput sequencing-ready mRNA, including DNA removal, small RNA and SSU rRNA removal, a higher volume of starting material was utilised for the extractions. As such, RNA was extracted from 5 g of colonised cotton, and 12 ml and 25 ml of concentrated soil-driven and biofilm-driven microcosm runoff, respectively. RNA yields were determined to be ~ 14 μg, ~ 4 μg and ~ 500 ng from the soil-driven microcosm, biofilm-driven microcosm and the colonised cotton, respectively. Following Turbo DNase treatment and clean-up using RNeasy MinElute kit, integrity of the RNA was determined using the prokaryote total RNA 6000 (nano) assay on a 2100 Bioanalyzer.

All 3 samples were determined to be degraded, as shown in the typical bioanalysis trace provided (Fig. 6.3). Contrary to the commonly observed RNA degradation, where numerous peaks of sheared SSU rRNA are found downstream of the 16S rRNA peak, a 'bell-shaped' distribution of RNA within the size range 800 bp-3 kb was observed. This was attributed to alkali-mediated RNA degradation, where the sample has the profile of an RNA sample that has been fragmented. As a result, it was concluded that neutralisation of the samples was probably required prior to cell lysis in order to protect the RNA from chemical scission.

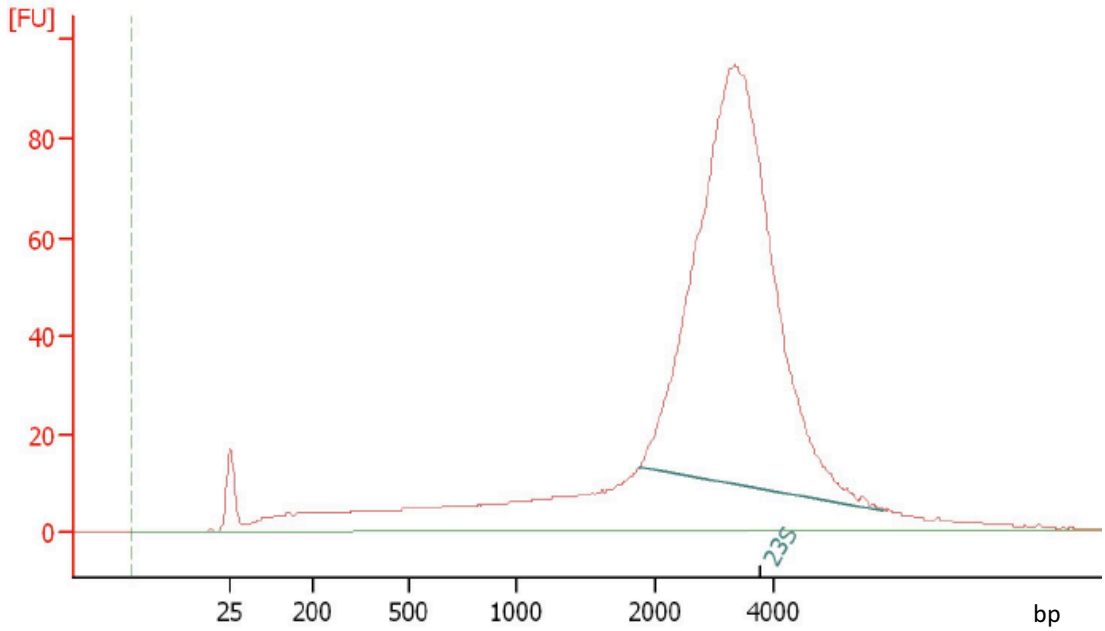


Figure 6.3 A typical bioanalyser trace of initial RNA extractions from the batch-fed microcosms or the colonised cotton incubated in the hyperalkaline site. A ‘bell-shaped’ distribution can be seen, with no SSU rRNA peaks visible.

6.4.2 Troubleshooting

Since high RNA yield was obtained from the soil-driven microcosm runoff, this was used to test neutralisation regimes to prevent the depletion of other precious samples. Various modifications were designed to potentially solve the issue, including:

- The pH of the CTAB buffer used in extraction was adjusted to pH 1.5 using concentrated HCl prior to extraction
- Acidic phenol:chloroform (5:1) (Sigma Aldrich) was used to perform extraction
- A combination of pH 1.5 CTAB and acidic phenol:chloroform (5:1) was used to perform extraction
- Extraction was performed with TRIzol reagent (Thermo Scientific) instead of phenol:choloform:isoamylalcohol (25:24:1)
- Extraction was performed using a combination of TRIzol reagent and pH 1.5 CTAB

No difference was observed in the RNA integrity profile using any of the above listed methods, i.e. the alkaline degradation of the RNA persisted.

6.4.3 RNA extraction from soil-leachate mixture

Due to the low yield observed from the colonised cotton and since a large quantity of DNA suitable for sequencing could be extracted from the anaerobically incubated, lignocellulose-rich soil-alkaline leachate mixture, RNA extraction was attempted on this material. Extraction of total community DNA and RNA from the same sample also aims to improve consistency, as direct comparison between the metagenomic and metatranscriptomic data can be made to draw conclusions on the functionality of the microbial community. The extraction was performed on 10 g of the soil-leachate mixture, with the addition of 2 x two volumes washes of the cellular pellet

using 1 x PBS (pH 4.0), as described in section 2.2.3. As with the DNA extract from the sample, high RNA yield of ~18 μg was obtained, and bioanalysis was performed following DNA removal and the subsequent clean-up (Fig. 6.4). Alkali-mediated RNA degradation was not observed, suggesting that the low pH buffer treatment was sufficient to neutralise the high pH of the leachate.

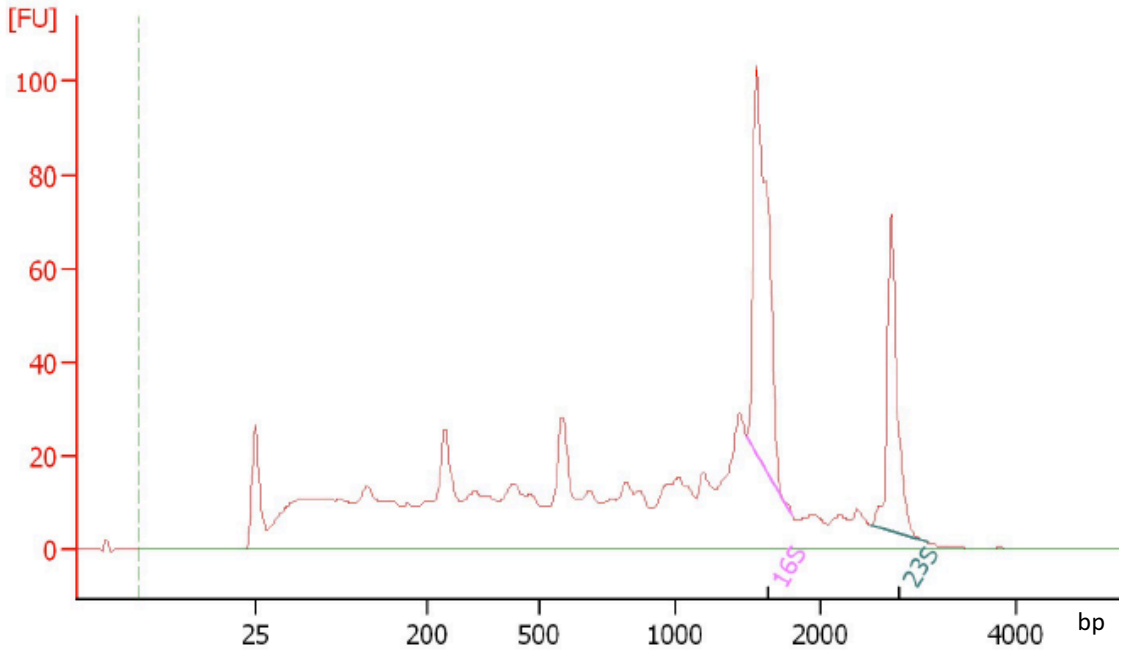


Figure 6.4 Bioanalysis of the RNA extracted from the soil-leachate mixture following low pH buffer washes. No alkali-mediated degradation can be seen, and the 16S and 23S rRNA peaks are clearly visible. Minor generic RNA degradation can be seen, however, as 2 small peaks at ca. 200 bp and 500 bp. The extracted RNA is still suitable for reverse transcription followed by high throughput sequencing of the cDNA generated.

6.4.4 RNA extraction from microcosm samples

The successful methodology was also tested on the run-off from the soil-driven microcosm, but the bioanalysis still indicated that chemical degradation was occurring. To troubleshoot, the low pH buffer wash was performed thrice in order to prevent RNA degradation. A similar RNA profile was observed, but the yield suffered (~ 5 μg) due to the loss of biomass as a result of the repeated washing-centrifugation cycle. In order to completely eradicate chemical RNA lysis without affecting the yield, particularly relevant for samples with low initial yield such as the biofilm-driven microcosm, a filter membrane-based method was developed.

Following differential centrifugation to remove the particulate matter, the cell suspension was filtered through a sterile, RNase-free 0.45 μm nitrocellulose membrane (PALL) to trap the cellular biomass. The cells were then washed using 200 ml of 1 x PBS (pH 4.0) whilst on the membrane, leading to a thorough neutralisation. RNA was subsequently extracted from the cells trapped on the membrane using the bead-beating method described by Griffiths *et al*. (2000) as described in section 2.2.2. RNA yields of ~ 20 μg and ~ 8 μg were obtained from the soil-based and biofilm-based microcosms, respectively, and the bioanalysis indicated that no alkali-mediated RNA degradation had occurred (Fig. 6.5).

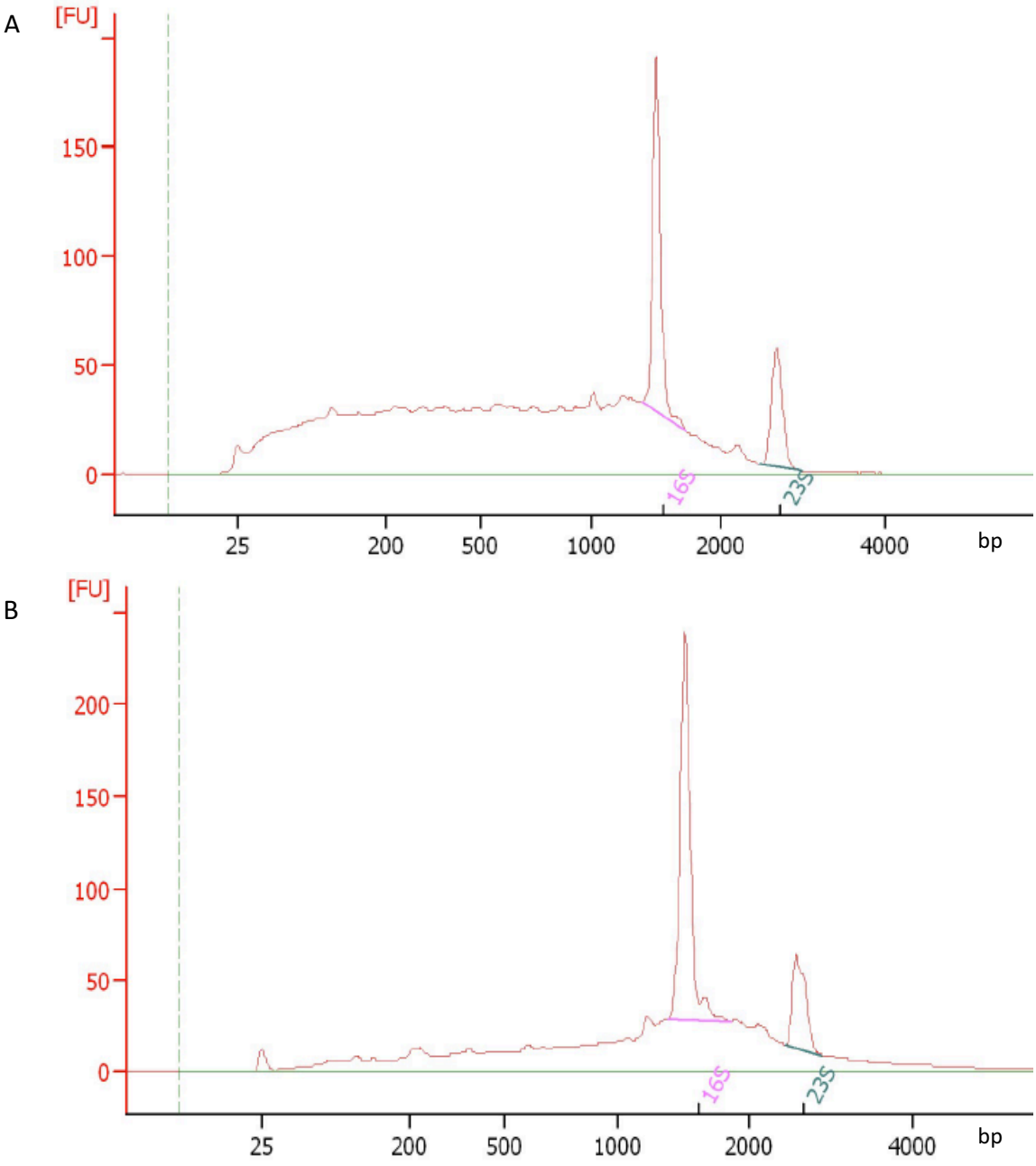


Figure 6.5 Bioanalysis of the RNA extracted from the soil-derived microcosm (A) and the biofilm-derived microcosm (B) following low pH buffer wash of cells trapped on a 0.45 μm membrane. No alkali-mediated degradation can be seen, and the 16S and 23S rRNA peaks are clearly visible.

6.5 Discussion

To summarise, community DNA and RNA were successfully extracted from the lignocellulose-rich soil-alkaline leachate mixture, as well as the run-off collected from the cotton biofilm-driven microcosm and the contaminated soil-driven microcosm. Metatranscriptomic RNA from all three samples will be reverse-transcribed into cDNA, followed by high throughput sequencing using the Illumina MiSeq platform at the Centre for Genomic Research, University of Liverpool. Mining of cellulases encoded in the soil-alkaline leachate metatranscriptome is particularly relevant to this project, whereas collaborators at the University of Huddersfield are interested in generating ISA metabolism pathways from the microcosm sequence data. Metagenomic DNA from the three samples can also be sequenced in a similar manner to enable comparative analysis of the genetic potential against the gene expression profile of the microbial communities (as presented in chapter 5).

Alkaliphilic bacteria are able to survive in high pH environments by maintaining their cytoplasmic pH 2.0-2.3 units below the external pH. These microbes are able to achieve this primarily through the action of Na+/H+ antiporters that use an electrochemical gradient of Na+, and proton-translocating ATP synthases for oxidative phosphorylation (Krulwich, 1995; Grant & Sorokin, 2011). Alkaliphilic cellulolytic bacteria have been well documented to grow at a pH of around pH 10.0 (Zhilina & Zavarzin, 1994; Zvereva *et al*., 2006) and although they have also been reported to be capable of growing at pH 12.0, there is some disagreement as to the accuracy of these reports (Sorokin, 2005). It has been suggested that there could be a limit as to the pH extreme at which these microbes can survive and grow, and this might indeed explain the paucity of biomass and the subsequent nucleic acids observed from the cotton

incubated in the hyperalkaline lagoon sediment. Indeed, a considerable difference was observed in nucleic acid yield between the colonised cotton and the soil-leachate mixture.

Such a difference in microbial biomass might not be unexpected, as the majority of the microbes would likely derive energy from the easier to assimilate and widely available ISA in the hyperalkaline sediment rather than the crystalline cellulose bait. It has been suggested that the microbial community at this particular site are able to degrade ISA at high pH (Burke *et al*., 2012; Rout *et al*., 2015), even when present as the sole carbon and energy source in minimal media (Bassil *et al*., 2014). The biofilm formation on the cotton was visibly sparse, yet the soil is full of microbes that can reportedly adapt to high pH. Burke *et al*. (2012) have also demonstrated that the soil at Buxton, where ISA is generated *in situ*, is microbially active despite pore waters of up to pH 13.0. Subsequent culturing of the alkaline sediment has established that the microbial communities inhabiting them are capable of α-ISA degradation under methanogenic conditions (Rout *et al*., 2015).

Studies under simulated repository conditions have estimated that chemical and radiolytic degradation accounts for up to 100% of the cellulose decomposition by one hundred thousand years post closure (Chambers *et al*., 2003) due to prevalence of conditions close to pH 12.0. Hence, it has been suggested that the vast majority of the cellulosic biomass will be degraded chemically before the pH recedes enough to allow for any significant microbial cellulose degradation, although CDPs present in the repository may be subjected to microbial degradation. A similar prospect might be observed in the hyperalkaline site where *in situ* generation of ISA occurs from the available cellulosic substrate, suggesting that identification of cellulases from transcriptomic data could prove to be challenging. However, microbial cellulose

degradation cannot be completely ruled out even at this high pH within repositories (Grant *et al*., 2001) and the hyperalkaline pond sediment due to the presence of suitable micro-niches that provide sufficient water and nutrient content. Cellulolytic metabolism might proceed in such niches due to microbial modification of local environmental conditions, given that anaerobic cellulose degradation generates organic acids that can help lower the pH (Chambers *et al*., 2004).

It is evident that DNA extraction from hyperalkaline environments is much more straightforward compared to RNA extraction. This is due to the difference in their chemical stability as a result of their different chemical structure. The presence of 2'-OH group in RNA renders it extremely susceptible to alkali-mediated scission even at pH 9.0, whereas DNA is much more stable at high pH. The double-stranded structure of the DNA can however, undergo denaturation and mild scission at extremely high pH (> pH 10.0). This might explain the relatively sheared nature of the extracted DNA in this study, as the extraction method used has previously yielded DNA predominantly in the size range ~ 10 kb.

The need for extra low pH buffer washes when extracting RNA from microcosm samples can be attributed to the fact that 4M NaOH was used to maintain the pH of 11.0 in the batch-fed systems. Even though the actual pH of the soil-leachate sample was higher (~ pH 13.0), the molarity of the microcosm samples was much higher, and hence required extensive neutralisation. The use of membrane entrapment method allows for extensive washing of the cells, whilst also ensuring minimal loss of biomass, leading to a higher RNA yield compared to neutralisation method using wash-centrifuge cycles. The difference in molarity of the alkali might also explain the slightly better DNA profile observed from the soil-leachate sample compared to the microcosm

samples, as a small fraction of the DNA could also be observed around the 10 kb-6 kb size range (Fig. 6.2).

The microbial community found within the cotton biofilm-driven was found to consist of flocculates, composed of bacteria embedded in polysaccharides and proteins stabilised by extracellular DNA (Charles *et al*., 2015). Although it might be possible that the DNA extraction method was unable to successfully dislodge the bacterial cells from the flocs, it is unlikely given that a similar method proved to be extremely potent at disrupting thick biofilm from colonised cotton and for nucleic acid extraction for the production of a fosmid library (chapter 4). As such, the difference in the yield between the soil-driven and the biofilm-driven microcosms is more likely to be due to the higher initial microbial load found associated with the soil in contrast to the cotton biofilm, leading to the subsequent development of a more populous microbial community in the former.

Chapter 7

General discussion

This thesis describes the results of an investigation of cellulases expressed by the microbial community present in anoxic landfill leachate by means of enrichment using a crystalline cellulose bait. It was hypothesised that specialist cellulose degrading microbes would colonise and decompose the cellulose bait, and metagenomic and metatranscriptomic analysis of such a community would demonstrate significant numbers of cellulolytic microbes as well as genes expressing glycoside hydrolases. High molecular weight metagenomic DNA from parallel samples was also used for the production of a fosmid library as functional screening of cloned DNA sequences in this manner can yield complete gene sequences encoding cellulases. Functional screening for the presence of genes encoding endoglucanases was performed on fosmid libraries generated from high molecular weight DNA extracted from cotton incubated in landfill leachate as well as lake sediment. Anaerobic cellulolytic fungi are known to be some of the most potent cellulose degraders in the microbial world, and an attempt was made to culture them from landfill leachate. Discovery of cellulases capable of functioning at extreme pH has industrial significance, and a method was developed for extracting high quality RNA from anaerobic hyperalkaline sediments (pH>12.0) to facilitate such discoveries using metatranscriptomic approaches. Some of the major conclusions from this study are as follows:

- PCR analysis confirmed the presence of cellulolytic clostridia and fibrobacters both in landfill leachate and dewaxed cotton incubated in the leachate, and less convincing evidence for the presence of amplified anaerobic fungal DNA was

also obtained. This was followed up with an unsuccessful attempt at culturing anaerobic chytrid fungi from landfill leachate.

- The identity of microbes producing a PCR positive result for 18S rRNA gene from leachate was queried by production and sequencing of a subset of a clone library, where ciliated protozoa of Class Armophorea accounted for the majority of the sequences. 454 pyrosequenced 18S rRNA gene amplicons from leachate and cotton incubated in leachate were however, discovered to be dominated by sequences that could only be assigned to unknown fungi.
- A metagenome, a metatranscriptome and a MessageAmp II aRNA amplification kit (Invitrogen) amplified metatranscriptome were derived from cotton incubated in landfill leachate and sequenced using Illumina MiSeq technology; the taxonomic analyses indicated that Bacteria were the dominant domain. However, a higher percentage of reads were classified as Eukaryota and Archaea in the metatranscriptome compared to the metagenome, suggesting that members of these domains were perhaps more active in the community than DNA analysis alone would suggest.
- There was a higher relative abundance of Euryarchaeota and unclassified reads in the metatranscriptome compared to the metagenome. The unclassified (derived from Eukaryota) reads almost entirely comprised of ciliated and flagellated protozoa, with members of Class Armophorea also present in higher numbers in the metatranscriptome than in the metagenome.
- Functional analyses of the datasets suggested that sequences annotated as having a function in clustering-based subsystems, carbohydrate metabolism, protein metabolism and RNA metabolism accounted for the vast majority of

reads assigned to the metagenome and the two metatranscriptomes, suggesting that the biofilm represented an active microbial community.

- No outstanding differences were evident in the taxonomic and functional comparison of the amplified and the non-amplified metatranscriptomes, suggesting that amplification using MessageAmp II aRNA amplification kit (Invitrogen) does not introduce significant bias to the microbial community profile.
- 1,724 ORFs predicted in metatranscriptomic data were assigned to glycoside hydrolase families by querying against Pfam-A database. Blastx analysis of these ORFs suggested that 114 sequences had no matches in NCBI nr database, and highest number of hits to a single ORF corresponded to a GH family 9 protein from an uncultured bacterium. The metagenome was also queried for sequences with 100% identity to ORFs corresponding to GH families, and some of the highly represented genes were annotated as putative or hypothetical protein. These results suggest that some novel GH enzymes are likely to be active in landfill leachate.
- A fosmid library derived from high molecular weight DNA extracted from cotton incubated in lake sediment was screened for expression of endoglucanases using the Congo red Assay, and two distinct clones were isolated and sequenced from a total of ~ 80,000. Assembly and analysis of one clone led to the identification of two genes encoding glycoside hydrolases.
- Another fosmid library was produced from high molecular weight DNA extracted from cotton incubated in landfill leachate, and preliminary screening was found to yield false positives.

- A method was developed to extract high quality total RNA, whilst maintaining sufficient yield, from mixed microbial communities found in anoxic hyperalkaline environments, including pH >12.0 sediment and pH 11.0 mesocosms.

Molecular microbial ecology studies focusing on environmental samples pose a number of issues. Sampling complex microbial communities from the environment is challenging, as samples potentially include archaeal, bacterial, eukaryotic and viral species at varying levels of diversity and abundance. Moreover, environmental microbial communities are highly heterogeneous and, as such, it is almost impossible to account for reproducibility within samples collected even within a few centimetres from each other. This, coupled with the frequent changes of conditions in the microenvironments over time and the response of the communities to such changes, leads to the formation of highly dynamic microbial populations that are constantly evolving, and their specific transcript content often displays temporal fluctuation (Uchiyama & Miyazaki, 2009; Gilbert *et al*., 2009; Andersson *et al*., 2010; Braga *et al*., 2016). Furthermore, the solid components of the leachate tend to settle at the bottom, as do some microbial species due to differences in physiological and behavioral lifestyle, leading to issues with reproducibility while sampling. These challenges are particularly applicable to leachate samples stored in carboys and may have had an impact on the composition of the microbial community that colonised and degraded the suspended dewaxed cotton, particularly as the DNA and RNA samples were not extracted concurrently. The metagenomic DNA used for Illumina sequencing was also extracted a few months prior to the high molecular weight DNA extracted for fosmid library production. Multiple replicates of the metagenome, the metatranscriptome and the amplified metatranscriptome are thus necessary to perform comprehensive

microbial community analyses complete with statistical analysis. As such, the comparison between microbial community metagenome, metatranscriptome and amplified metatranscriptome presented here should be considered with a degree of caution.

The establishment of a 'rare biosphere' has been widely reported in multiple environments (Lynch & Neufeld, 2015), and is equally applicable to landfill leachate. While the majority of the species in such an environment are observed to be low in abundance, it has been reported that the abundance of such rare species can fluctuate by 2-3 orders of magnitude depending on the time of sampling (Andersson *et al*., 2010). Furthermore, Galand *et al*. (2009) demonstrated that samples collected in such environments showed greater resemblance in terms of community composition when collected from microenvironments with similar physiochemical properties rather than when collected from microenvironments that were geographically closer. This can be attributed to the multitude of species found at a low abundance adapting specifically to specialised microenvironments, making environmental sampling even more challenging. Dao *et al*. (2016) reported seasonal variation in microbial community composition in landfill leachate. High throughput sequencing of microbial community 16S rRNA gene using the Illumina MiSeq pointed towards a difference in the predominant Bacterial and Archaeal Orders between rainy and dry seasons in a leachate treatment system.

High throughput sequencing platforms require reverse-transcribed cDNA as the template, generation of which poses issues of its own. Increase in length of RNA transcripts is inversely correlated to efficiency of reverse-transcription, as reverse transcriptases might introduce errors during cDNA synthesis (Roberts *et al*., 1989; Stewart *et al*., 2010). Additionally, presence of high homology regions in the cDNA

can lead to the generation of chimeras due to template switching by reverse transcriptases (Cocquet *et al*., 2006). This was perhaps evident in the metatranscriptomic datasets, as a high percentage of reads were determined to be artificial replicates related to SSU and LSU rRNA sequences.

Despite the aforementioned issues, the advent of high throughput sequencing now allows us to not only sequence individual genomes completely, but metagenomes too with sufficient depth and coverage to facilitate detection of rare and less abundant members of microbial communities. The generation of a vast amount of metatranscriptomic data enables characterisation of transcripts circumventing the need for any prior knowledge of their nucleotide sequence (Carvalhais *et al*., 2012). The steady increase in throughput and decrease in sequencing costs, along with improvements in our ability to analyse and annotate huge sequence datasets is transforming the application of these 'omics' technologies to the field of molecular microbial ecology (Scholz *et al*., 2012).

Given that the majority of our understanding of microbial cellulose decomposition stems from the herbivore gut, this study has built on our existing knowledge of the occurrence of this important ecological process in other anoxic environments. Lignocellulosic biomass comprises a large proportion of biodegradable matter in landfill sites, and anaerobic degradation of cellulose in this environment was originally thought to have been mediated predominantly by members of Firmicutes class Clostridia and some Bacteroidetes (Leschine, 1995; Burrell *et al*., 2004). However, there is increasing evidence that distribution of cellulolytic microbes such as Fibrobacteres and anaerobic chytrid fungi is not restricted to the rumen (Ransom-Jones *et al*., 2012; McDonald *et al*., 2012). This study indicates that these microbes, along with *Sorangium* and anaerobic ciliated protozoa belonging to the class Armophorea are

distributed in landfill leachate, and are likely to be involved in cellulose decomposition in this environment.

Ciliates in class Armophorea are particularly intriguing, as members have been reported to be involved in cellulose degradation in wood-feeding cockroach and termite gut (Gijzen *et al*., 1994; Lynn & Wright, 2013), but no literature exists addressing their abundance in landfill leachate. Their classification under Unclassified (derived from Eukaryota) by MG-RAST also points towards little understanding of their distribution and ecological importance. It is also interesting that ciliates in the rumen with a defined role in lignocellulose decomposition, including *Entodinium*, *Epidinium* and *Polyplastron*, are classified under class Litostomatea and not Armophorea (Wright *et al*., 1997). It is possible that members of class Armophorea are present in other anoxic environments also and design of specific primers could allow for PCR-mediated analysis of their environmental distribution, potentially allowing for elucidation of their role in cellulose degradation.

Data collected from 454 pyrosequencing of 18S ribosomal DNA amplicons suggested that fungi dominate the eukaryotic component of the microbial community in landfill leachate as well as in the biofilm colonising the cotton incubated in leachate. However, this is inconsistent with the rest of the data generated in this study, as PCR amplification of 18S rRNA gene as well as metagenomic and metatranscriptomic analyses indicated that anaerobic fungi are present in much smaller, yet environmentally significant numbers in the leachate samples; the former result is likely to be a product of temporal variation. Furthermore, their distribution has been determined to be highly sporadic in landfill leachate samples, as leachate in the Bidston Moss site and only riser 2 in the Bromborough Dock site have produced molecular ecological evidence for their presence (McDonald *et al*., 2012). It is also worth pointing

out that representatives of cellulolytic microbes for which PCR analysis was performed initially, clostridia and fibrobacters, were found to be present in both the metagenomic as well as the metatranscriptomic datasets, with representatives of clostridia much more abundant than those of fibrobacters.

'Omics' approaches are particularly favourable for gene mining when compared to conventional molecular techniques such as PCR, as designing specific primers based on sequences present in the database is required for the latter, meaning identification of truly novel and undiscovered genes is unattainable (Smith & Osbourne, 2009). High throughput sequencing has the potential to detect hundreds of interesting enzymes (glycoside hydrolases in the case of this study), and coupling of this approach with fosmid-based clone library screening can lead to the isolation of whole genes for enzyme overexpression and characterisation (Geng *et al*., 2012; Xia *et al*., 2013; Rooks *et al*., 2012). Such an approach has been used for identification of carbohydrate active enzymes in metagenomes (Rashamuse *et al*., 2013), and to a much lesser extent coupled with metatranscriptomics (Zifcakova & Baldrian, 2012). With the ever-increasing throughput and decreasing costs of sequencing, the major bottleneck to analysis and annotation of several hundred Gigabases of sequencing data is now of a computational nature (Stein, 2010; Abram, 2015).

Although high throughput sequencing only generates *in silico* data, it provides a rich repertoire of sequences from uncultivable microbes that can be used to examine the metagenome for enzymes of interest and their subsequent overexpression following molecular cloning (Li *et al*., 2009). Lehembre *et al*. (2013) employed a functional metatranscriptomic approach for the analysis of genes responsible for heavy metal resistance in soil microbial communities. The reverse transcription and subsequent cloning and screening of Poly-A tailed mRNA led to the identification of novel

eukaryotic genes with no database homologues that were responsible for heavy metal resistance. This method offers an alternative approach to the methodology suggested here, where the generated metatranscriptomic data was used to identify novel transcripts with cellulolytic function for the potential design of PCR primers and molecular probes to enable sequence-based screening of the fosmid libraries generated from the same microbial community. This is particularly useful for identification of eukaryotic sequences, given the highly cellulolytic nature of fungi *Neocallimastigales*, as it is not possible to successfully screen eukaryotic enzymes during functional screening of *E. coli* fosmid libraries due to the inherent differences in prokaryotic and eukaryotic translational machineries as well as post-translational modification and secretion pathways.

It is also important that attempts to culture microbes with impressive cellulolytic capability are continued from environments such as landfill leachate, as an increasing amount of data generated through high throughput sequencing of environmental communities corresponds to hypothetical proteins originating from uncultured organisms. Culturing and subsequent isolation of anaerobic microbes such as cellulolytic fungi will allow for genome sequencing of close relatives of fungi that display outstanding ability to decompose cellulose in the herbivore gut (Youssef *et al.*, 2013). This will facilitate discovery of potentially novel full length genes through genome assembly, while metatranscriptomic analysis might just have indicated a putative function for only a small section of the same gene.

To date, there have been no publications focussing on the microbial ecology of cellulose degradation in landfill leachate that have combined metagenomic and metatranscriptomic sequencing of the microbial community with a fosmid library screening approach, making the investigation detailed in this study novel and exciting.

Moreover, the metagenome, the metatranscriptome and the fosmid library generated from the colonised cotton bait incubated in landfill leachate represent three massive datasets which could be used for the identification of all manner of hydrolytic enzymes. Identification of 1,724 genes representing glycoside hydrolases from the metatranscriptome suggests that lignocellulolytic enzymes other than those involved in cellulose hydrolysis, such as xylanases, can also be sought from the datasets.

Mounting concerns over greenhouse gas emission and an increased demand for renewable energy renders the search for novel cellulases with characteristics applicable in industry vital. Examination of cellulolytic microbial communities that have evolved independently in environments under tough selection pressures, such as in anoxic landfill leachate or alkaline anoxic sediment, can lead to the discovery of enzymes with high specific activity for cellulose decomposition even in extreme industrial conditions. Production of second generation bioethanol is currently not cost-effective, and enzymes able to catalyse the hydrolysis of lignocellulosic biomass into fermentable sugars at an elevated rate can dramatically reduce the cost of production through their inclusion in consolidated biomass processing (Brethauer & Studer, 2015).

Molecular microbial ecology is centered on understanding the community structure, function and fitness of microbial populations in response to change in conditions, and as such, should be studied over a period of time to reflect the metabolic state of the microbes, along with extensive replicates and adequate controls. This is particularly true for metatranscriptomics, as such an analysis is largely dependent on environmental conditions at a given moment and the effect it has on community gene expression in response to those specific conditions. However, the financial constraints with high-throughput sequencing and the complexity of bioinformatic analyses of

multiple environmental samples means that such a thorough investigation was beyond the scope of this study.

While gene mining based on metatranscriptomic approaches is relatively challenging, given the small proportion of mRNA present in a microbial community total RNA extract, an increasing number of studies are emerging that involve high throughput sequencing of community RNA. Ribo-Zero rRNA depletion kit (Epicentre) has been reported to be highly efficient at rRNA sequence removal from community total RNA extracts (Giannoukos *et al*., 2012), while rRNAFilter (Wang *et al*., 2016) has been developed recently as a tool for quick removal of rRNA reads without the requirement for a reference database. To that effect, rRNA sequence removal has been reported to be more comprehensive than SortMeRNA or RiboPicker using this tool. Only nanogram amounts of mRNA enriched/rRNA depleted total RNA extract is required for sequencing applications, as seen in ScriptSeq and NextEra library preparation kits for Illumina MiSeq, to draw accurate biological conclusions from the data output.

It is possible that single cell technologies might replace metagenomic and metatranscriptomic studies that are a popular method of microbial community analysis at the time of writing. The ability to use single cell genomic and transcriptomics for studying uncultivable cells from environmental samples bypasses some of the difficulties associated with analysis of large metagenomic and metatranscriptomic datasets by reducing sample heterogeneity (Rinke *et al*., 2013; Kodzius & Gojobori, 2016). For instance, sequencing data from a single organism is much simpler to assemble and analyse using Blast, leading to a more comprehensive and detailed investigation into gene expression. Moreover, coverage is no longer an issue as it is in

metagenomic studies, as the latter only generates a partial overview of complex microbial communities (Liang *et al*., 2014).

In the near future, it is also likely that an integrated approach involving metagenomics, metatranscriptomics, metaproteomics, metabolomics as well as stable isotope probing (SIP) will lead the way in molecular microbial ecology (Abram, 2015). In-depth analysis of microbial community structure, function, activity as well as interaction using these systems-based approaches will be possible even from highly complex environmental samples (Aguiar-Pulido *et al*., 2016). Given however the pace at which sequencing technology is evolving, it is not unlikely that a completely new workflow for gene and enzyme discovery as well as for defining ecological roles of specific microbial communities will have evolved in the next decade, especially from never studied before environments.

References

Aakvik, T., Degnes, K. F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., Volker, U., Ellingsen, T. E. & Valla, S. (2009). A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbial Lett*. **296(2)**, 149-158.

Abram, F. (2015). Systems-based approaches to unravel multi-species microbial community functioning. *Comput. Struct. Biotechnol. J*. **13**, 24-32.

Affa'a, F. M., Ndongo, N., & Granosik, L. (1995). Étude morphologique et morphométrique des ciliés Prosicuophora , commensaux de Bufo regularis et B. maculatus de la région de Yaoundé (Cameroun). *Archiv für Protistenkunde* **145**, 127–131.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D. *et al*. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. **44(W1)**, W3-W10.

Agatha, S., Strüder−Kypke, M. C., Beran, A., & Lynn, D. H. (2005). Pelagostrobilidium neptuni (Montagnes and Taylor, 1994) and Strombidium biarmatum nov. spec. (Ciliophora, Oligotrichea): Phylogenetic position inferred from morphology, ontogenesis, and gene sequence data. *Eur. J. Protistol.* **41**, 65–83.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K. & Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evol. Bioinform. Online* **12(12)**, 5-16.

Aird, D., Ross, M. G., Chen, W., Danielsson, M., Fennell, T., Russ, C., Jaffe, D., Nusbaum, C. & Gnirke, A. (2011). Analysing and minimising PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, 18-31.

Akin, D. E. & Rigsby, L. L. (1987). Mixed fungal populations and lignocellulosic tissue degradation in the bovine rumen. *Appl. Environ. Microbiol.* **53**, 1987-1995.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol*. **215**,403-410.

Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143-169.

Andersson, A. F., Riemann, L. & Bertilsson, S. (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME Journal* **4**, 171-181.

Ando, S., Ishida, H., Kosugi, Y. & Ishikawa, K. (2002). Hyperthermostable endoglucanases from *Pyrococcus horikoshii*. *Appl. Environ. Microbiol.* **68**, 430-433.

Artzi, L., Bayer, E. A. & Morais, S. (2017). Cellulosomes: Bacterial nanomachines for dismantling plant polysaccharides. *Nature Rev. Microbiol.* **15**, 83-95.

Ball, A. S., Brabban, A, & McCarthy, A. J. (1992). Studies on the extracellular p-nitrophenyl b-D-cellobiosidase activity of a thermophilic streptomycete. *Appl. Microbiol. Biotechnol*. **36**, 473-477.

Baldrian, P., Valaskova, V., Merhautova, V. & Gabriel, J. (2005). Degradation of lignocellulose by *Pleurotus ostreatus* in the presence of copper, manganese, lead and zinc. *Res. Microbiol*. **156**, 670-676.

Balzer, S., Malde. K., Lanzen, A., Sharma, A. & Jonassen, I. (2010). Characteristics of 454 pyrosequencing data- enabling realistic simulation with flowsim. *Bioinformatics* **26(18)**, 420-425.

Bao, L., Huang, Q., Chang, L., Zhou, J. & Lu, H. (2011). Screening and characterization of a cellulase with endoglucanase and exocellulases activity from yak rumen metagenome. *J. Mol. Catal. B: Enzym*. **73**, 104-110.

Barlaz, M. A. (1996). Microbiology of solid waste landfills. In *Microbiology of solid waste*. Palmisano, A. C. and Barlaz, M. A. (ed): CRC press.

Barlaz, M. A. (1997). Microbial studies of landfills and anaerobic refuse decomposition. In *Manual for Environmental Microbiology*. Washington, D. C.: American Society for Microbiology.

Barlaz, M. A. (2006). Forest products decomposition in municipal solid waste landfills. *Waste Management* **26**, 321-333.

Barlaz, M. A., Schaefer, D. M. & Ham, R. K. (1989). Bacterial population development and chemical characteristics of refuse decomposition in a simulated sanitary landfill. *Appl. Environ. Microbiol.* **55**, 55-65.

Barlaz, M. A. & Ham, R. K. (1993). Leachate and gas generation. *In Geotechnical practice for waste disposal*. (pp 113-136). Chapham & Hall, London.

Barlaz, M. A., Kaplan, P. O., Ranjithan, S R. & Rynk, R. (2003a). Comparing recycling, composting and landfills. *Biocycle* **44**, 60-64.

Barlaz, M. A., Kaplan, P. O., Ranjithan, S R. & Rynk, R. (2003b). Evaluating environmental impacts of solid waste management activities. *Biocycle* **44**, 52-56.

Barley, E. & Fitzpatrick, K. (2011). The structure and function of large biological molecules. In *Lecture presentations for Campbell Biology*. Pearson Education Inc. Ninth ed. 2011.

Bashiardes, S., Zilberman-Schapira, G. & Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* **10**, 19-25.

Bassil, N. M., Bryan, N. & Lloyd, J. R. (2014). Microbial degradation of isosaccharinic acid at high pH. *The ISME Journal,* 1–11.

Bauchop, T. (1979). Rumen anaerobic fungi of cattle and sheep. *Appl. Environ. Microbiol.* **38**, 148-158.

Bauchop, T. (1981). The anaerobic fungi in rumen fiber digestion. *Agriculture and Environment* **6**, 339-348.

Bayer, E. A., Belaich, J. P., Shoham, Y. & Lamed, R. (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol*. **58**, 521-554.

Bayer, E. A., Lamed, R., White, B. A. & Flint, H. J. (2008). From cellulosomes to cellulosomics. *The Chemical Record* **8**, 364-377.

Bayer, T. S., Widmaier, D. M., Temme, K., Mirsky, E. A., Santi, D. V. & Voigt, C. A. (2009). Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc*. **131(18)**, 6508-6515.

Beeson, W. T., Vu, V. V., Span, E. A., Phillips, C. M. & Marletta, M. A. (2015). Cellulose degradation by polysaccharide monooxygenases. *Annu. Rev. Biochem*. **84**, 923-946.

Beguin, P. & Aubert, J. (1994). The biological degradation of cellulose. *FEMS Microbiol. Rev.* **13(1)**, 25-58.

Bekele, A. Z., Koike, S., & Kobayashi, Y. (2011). Phylogenetic diversity and dietary association of rumen Treponema revealed using group-specific 16S rRNA gene-based analysis. *FEMS Microbiol. Lett.*, **316(1),** 51-60.

Berini, F., Presti, I., Beltrametti, F., Pedroli, M., Varum, K. M., Pollegioni, L., Sjoling, S. & Marinelli, F. (2017). Production and characterization of a novel antifungal chitinase identified by functional screening of a suppressive-soil metagenome. *Microb. Cell Fact*. **16(1)**, 16.

Bernstein, J. A., Khodursky, A. B., Lin, P., Lin-Chao, S. & Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-colour fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U. S. A.* **99(15)**, 9697-9702.

Bernstein J. R., Butler T., Shen C. R. & Liao J. C. (2007). Directed evolution of ribosomal protein S1 for enhanced translational efficienty of high CG Rhodopseudomonas palustris DNA in *Escherichia coli. J. Biol. Chem.* **282(26)**,18929–18936.

Billon-Grand, G., fiol, J. B., Breton, A., Bruyere, A. & Oulhaj, Z. (1991). DNA of some anaerobic rumen fungi- G+C content determination. *FEMS Microbiol. Lett.* **82**, 267-270.

Binga, E. K., Lasken, R. S., & Neufeld, J. D. (2008). Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. *The ISME Journal* **2(3),** 233-241.

Birbir, M. Calli, B., Mertoglu, B., Bardavid, R. E., Oren, A., Ogmen, M. N. & Ogan, A. (2007). Extremely halophilic Archaea from Tuz Lake, Turkey, and the adjacent Kaldirim and Kayacik salterns. *World J. Microbiol Biotechnol.* **23**, 309-316.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., & Nekrutenko, A. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics*, ***26*(14),** 1783-1785.

Bonhomme, A. (1990). Rumen ciliates: Their metabolism and relationships with bacteria and their hosts. *Anim. Feed Sci. Technol.* **30**, 203–266.

Bonhomme−Florentin, A. (1994). Ciliés du rumen – métabolisme et relations avec les bactéries et leur hôte. In P. de Puytorac (Ed.), Traité de zoologie, infusoires ciliés (Vol. 2, pp. 381–394). Paris: Masson.

Borneman, S. & Akin, D. E. (1994). The nature of anaerobic fungi and their polysaccharide degrading enzymes. *Mycoscience* **35**, 199-211.

Bove, R. & Lunghi, P. (2006). Electric power generation from landfill as using traditional and innovative technologies. *Energy Conservation and Management* **47**, 1391-1401.

Braga, R. M., Dourado, M. N. & Araujo, W. L. (2016). Microbial interactions: Ecology in a molecular perspective. *Braz. J. Microbiol*. **47 (1)**, 86-98.

Bremer, K. (1985). Summary of green plant phylogeny and classification. *Cladistics* **1(4)**, 369-385.

Brennerova, M. V., Josefiova, J., Brenner, V., Pieper, D. H. & Junca, H. (2009). Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities rom soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ. Microbiol*. **11(9)**, 2216-2227.

Brethauer, S. & Studer, M H. (2015). Biochemical conversion processes of lignocellulosic biomass to fuels and chemicals - A Review. *Chimia (Aarau)* **69(10)**, 572-581.

Brock, T. D., Madigan, M. T., Martinko, J. M. & Parker, J. (1994). *Biology of Microorganisms*, 7th ed., Prentice Hall, Englewood Cliffs, NJ. p. 650.

Brookman, J. L., Mennim, G., Trinci, A. P. J., Theodorou, M. K. & Tuckwell, D. S. (2000a). Identification and characterisation of anaerobic gut fungi using molecular

methodologies based on ribosomal ITS1 and 18S rRNA. *Microbiology* **146(2)**, 393-403.

Brookman, J. L., Ozkose, E., Rogers, S., Trinci, A. P. J. & Theodorou, M. K. (2000b). Identification of spores in the polycentric anaerobic gut fungi which enhance their ability to survive. *FEMS Microbiol Ecol*. **31(3)**, 261-267.

Brownlee, A. G. (1988). A rapid DNA isolation procedure applicable to any refractory filamentous fungi. *Fungal Genet. Newsl.* **35**, 8.

Brownlee, A. G. (1989). Remarkably AT-rich genomic DNA from the anaerobic fungus *Neocallimastix*. *Nucleic Acids Res*. **17(4)**, 1327-1335.

Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., Coutinho, P. M., Henrissat, B., Nelson, K. E. & White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1948-1953.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. *Nature Rev. Microbiol*. **12**, 168-180.

Bryant, M. P. (1952). The isolation and characteristics of a spirochete from the bovine rumen. *J.Bact.* ***64*****(3),** 325.

B.S.I. BS ISO 14853:2005 Plastics-Determination of the ultimate anaerobic biodegradation of plastic materials in an aqueous system- Method by measurement of biogas production. British Standards Institute, London, UK. 2005.

Buitkamp, U. (1977). Die Ciliatenfauna der Savanne von Lamto (Elfenbeinküste). *Acta Protozoologica* **16** , 249–276.

Buranov, A. U. & Mazza, G. (2008). Lignin in straw of herbaceous crops. *Ind. Crop Prod*. **28**, 237-259.

Burke, I. T., Mortimer, R. J. G., Palani, S., Whittleston, R. A., Lockwood, C. L., Ashley, D. J. & Stewart, D. I. (2012). Biogeochemical reduction processes in a hyperalkaline leachate affected soil profile. *Geomicrobiol. J.* **29(9)**, 769-779.

Burnley, S. J. (2007). A review of municipal waste composition in the United Kingdom. *Waste Management* **27**, 1274-1285.

Burrell, P. C., O'Sullivan, C., Song, H., Clarke, W. P. & Blackall, L. L. (2004). Identification, detection, and spatial resolution of Clostridium populations responsible for cellulose degradation in a methanogenic landfill leachate bioreactor. *Appl. Environ. Microbiol*. **70(4)**, 2414-2419.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) "BLAST+: architecture and applications." *BMC Bioinformatics* **10**, 421.

Cameron, S. L., & O'Donoghue, P. J. (2003b). Trichostome ciliates from Australian marsupials. IV. Distribution of the ciliate fauna. *Eur. J. Protistol.* **39**, 139–147.

Cantarel B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenommics. *Nucleic Acids Res.* **37**, 233-238.

Cantarel, B. L., Lombard, V. & Henrissat, B. (2012). Complex carbohydrate utilization by the healthy human microbiome. *PloS One* **7**, e28742.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K. *et al*. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7(5)**, 335-336.

Carl, G. R., & Brown, R. D. (1983). Protozoa in the forestomach of the collared peccary (*Tayassu tajacu*). *J. Mammal.* **64**, 709.

Carvalhais, L. C., Dennis, P. G., Tyson, G. W. & Schenk, P. M. (2012). Application of metatranscriptomics to soil environments. *J. Microbiol. Meth.* **91**, 246-251.

Chambers, A. V., Gould, L. J., Harris, A. W., Pilkington, N. J. & Williams, S. J. (2003). Evolution of the near field of the nirex deposit concept. AEA Technology Report AEAT/R/ENV/0236.

Chambers, A. V., Holtom, G. J., Hunter, F. M. I., Ilett, D. J., Tearle, W. M. & Williams, S. J. (2004). pH evolution in super compacted waste. Serco Assurance Report SERCO/ERRA-0444.

Charles, C., Rout, S., Garratt, E., Patel, K., Laws, A. and Humphreys, P. (2015). The enrichment of an alkaliphilic biofilm consortia capable of the anaerobic degradation of isosaccharinic acid from cellulosic materials incubated within an anthropogenic, hyperalkaline environment. *FEMS Microbiol. Ecol.* . ISSN 1574-6941.

Chen, H. L., Chen, Y. C., Lu, M. Y., Chang, J. J., Wang, H. T., Ke, H. M. *et al*. (2012a). A highly efficient beta-glucosidase from the buffalo rumen fungus *Neocallimastix patriciarum* W5. *Biotechnol. Biofuels* **5(1)**, 24.

Chomczynski, P. & Sacchi, N. (1987). Single-step method of RNS isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem*. **162(1)**, 156-159.

Chundawat, S. P. S., Beckham, G. T., Himmel, M. E. & Dale, B. E. (2011). Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu. Rev. Chem. Biomol. Eng*. **2(6)**, 1-25.

Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88(1)**, 127-131.

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Bandela, A. M., Cardenas, E., Garrity, G. M. & Tiedje, J. M. (2007). The

ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. *Nucleic Acids Res*. **35**, 169-172.

Collins, M. D., Lawson, P. A., Willems, A., Cordoba, J. J., Fernandezgarayzabal, J., Garcia, P., Cai, J., Hippe, H. & Farrow, J. A. E. (1994). The Phylogeny of the genus *Clostridium* - Proposal of 5 new genera and 11 new species combinations. *Int. J. Syst. Bacteriol.* **44**, 812-826.

Comlekcioglu, U., Ozkose, E., Tutus, A., Akyol, I. & Ekinci, M. S. (2010). Cloning and characterisation of cellulase and xylanase coding genes from anaerobic fungus *Neocallimastix* sp GMLF1. *Int. J. Agric. Biol.* **12(5)**, 691-696.

Cottrell, M. T., Yu, L. & Kirchman, D. L. (2005). Sequence and expression analyses of Cytophaga-like hydrolases in a Western arctic metagenomic library and the Sargasso Sea. *Appl. Environ. Microbiol.* **71 (12)**, 8506-8513.

Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C. & Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse Proteobacteria. *Appl. Environ. Microbiol*. **76(5)**, 1633-1641.

Curtis, T. P. & Sloan, W. T. (2004). Prokaryotic diversity and its limits: Microbial community structure in nature and implications for microbial ecology. *Curr. Opin. Microbiol*. **7(3)**, 221-226.

Dabney, J. & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* **52**, 87-94.

Dao, H. T., Kuroda, K., Nakahara, N., Danshita, T., Hatamoto, M. & Yamaguchi, T. (2016). 16S rRNA gene-based comprehensive analysis of microbial community compositions in a full-scale leachate treatment system. *J. Biosci. Bioeng*. **122(6)**, 708-715.

Daly, K., Sharp, R. J. & McCarthy, A. J. (2000). Development of oligonucleotide probes and PCR primers for detecting phylogenetic subgroups of sulphate-reducing bacteria. *Microbiology UK* **146**, 1693-1705.

Daniel, D. E. (1993). *Geotechnical practice for waste disposal*. London: Chapman and Hall.

Davenport, C. F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S. *et al*. (2012). Genometa- A fast and accurate classifier for short metagenomic shotgun reads. *PLoS One* **7(8)**, e41224.

Davies, D. R., Theodorou, M. K., Lawrence, M. I. & Trinci, A. P. (1993). Distribution of anaerobic fungi in the digestive tract of cattle and their survival in faeces. *J. Gen. Microbiol*. **139(6)**, 1395-1400.

Dawson, S. C. & Pace, N. R. (2002). Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8324-8329.

De Filippo, C., Ramazzotti, M., Fontana, P. & Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief. Bioinform.* **13**, 696-710.

de Menezes, A. B., Lockhart, R. J., Cox, M. J., Allison, H. E. & McCarthy, A. J. (2008). Cellulose degradation by Micromonosporas recovered from freshwater lakes and classification of these Actinomycetes by DNA gyrase B gene sequencing. *Appl. Env. Microbiol.* **74**, 7080-7084.

de Menezes, A. B., McDonald, J. E., Allison, H. E. & McCarthy, A. J. (2012). Importance of Micromonospora spp. as colonizers of cellulose in freshwater lakes as demonstrated by quantitative reverse transcriptase PCR of 16S rRNA. *Appl. Environ. Microbiol.* **78**, 3495-3499.

Dehority, B. A. (1995). Rumen ciliates of the pronghorn antelope (*Antilocapra americana*), mule deer (*Odocoileus hemionus*), white−tailed deer (*Odocoileus virginianus*) and elk (*Cervus canadensis*) in the northwestern United States. *Archiv für Protistenkunde* **146**, 29–36.

Dehority, B. A. (2003). *Rumen Microbiology*. Nottingham, UK: Nottingham University Press.

Delmont T. O., Robe P., Clark I., Simonet P. & Vogel T. M. (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol. Methods* **86(3)** 397–400.

Demain, A. L., Newcomb, M. & Wu, J. H. (2005). Cellulase, clostridia and ethanol. *Microbiol. Mol. Biol. Rev.* **69**, 124-154.

Demeyer, D. I. (1981). Rumen microbes and digestion of plant cell walls. *Agricult. Environ.* **6(2)**, 295-337.

Deng, W., Xi, D., Mao, H. & Wanapat, M. (2008). The use of molecular techniques based on ribosomal RNA and DNA for rumen microbial ecosystem studies: A review. *Mol. Biol. Rep.* **35**, 265-274.

Denman, S. E. & McSweeney, C. S. (2006). Development of a real-time PCR assay for monitoring anaerobic fungal and cellulolytic bacterial populations within the rumen. *FEMS Microbiol. Ecol.* **58**, 572-582.

Department for Environment, Food and Rural Affairs (2009). Municipal waste composition: a review of municipal waste component analyses. *DEFRA*, UK.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol*. **72(7)**, 5069-5072.

Deshpande, M., V., Peterson, L., G. & Eriksson, L. G. (1988). A selective assay for the exo 1,4 b-glucanases. *Methods Enzymol*. **160**, 126-130.

Deutscher, D., Meilijson, I., Kupiec, M & Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet*. **38(9)**, 993-998.

Devillard, E., Newbold, C. J., Scott, K. P., Forano, E., Wallace, R. J., Jouany, J. P., & Flint, H. J. (1999). A xylanase produced by the rumen anaerobic protozoan *Polyplastron multivesiculatum* shows close sequence similarity to family 11 xylanases from gram-positive bacteria. *FEMS Microbiol. Lett.* **181(1)**, 145-152.

Ding, S. Y., Rincon, M. T., Lamed, R., Martin, J. C., McCrae, S. I., Aurilia, V. *et al*. (2001). Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. *J. Bact.* **183**, 1945-1953.

Duan, C. J., Xian, L., Zhao, G. C., Feng, Y., Pang, H. E., Bai, X. L. *et al*. (2009). Isolation and partial characterisation of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J. Appl. Microbiol.* **107(1)**, 245-256.

Durand, R., Rascle, C., Fischer, M. & Fevre, M. (1997). Transient expression of the beta-glucuronidase gene after biolistic transformation of the anaerobic fungus *Neocallimastix frontalis*. *Curr. Genet*. **31(2)**, 158-161.

Eadie, J. M., & Gill, J. C. (1971). The effect of the absence of rumen ciliate protozoa on growing lambs fed on a roughage–concentrate diet. *British Journal of Nutrition* **26(2)**, 155-167.

Edwards, U., Rogall, T., Blocker, H., Emde, M. & Bottger, E. C. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucl. Acids Res.* **17**, 7843-7853.

Edwards, J. L., Smith, D. L., Connolly, J., McDonald, J. E., Cox, M. J., Joint, I., Edwards, C. & McCarthy, A. J. (2010). Identification of carbohydrate metabolism genes in the metagenome of a marine biofilm community shown to be dominated by Gammaproteobacteria and Bacteroidetes. *Genes* **1**, 371-384.

Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & van Elsas, J. D. (2012). The great screen anomaly- a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol*. **93**, 1005-1020.

Engelbrektson, A., Kunin, V., Wrighton, K. C., Zvenigorodsky, N., Chen, F., Ochman, H. & Hugenholtz, P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* **4**, 642-647.

Environment Agency (2006). Waste data update 2006. Environment Agency.

Farquhar, G. J. & Rovers, F. A. (1973). Gas production during refuse decomposition. *Water, Air and Soil Pollution* **2**, 483-495.

Feldman AJ, Costouros NG, Wang E, Qian M, Marincola FM, Alexander HR, & Libutti SK (2002). Advantages of mRNA amplification for microarray analysis. *Biotechniques* **33(4)**, 906–914.

Fenchel, T. (1993). Methanogenesis in marine shallow water sediments: The quantitative role of anaerobic protozoa with endosymbiotic methanogenic bacteria. *Ophelia* **37**, 67–82.

Ferrer, M., Chernikova, T. N., Timmis, K. N. & Golyshin, P. N. (2004). Expression of a temperature-sensitive esterase in a novel chaperone-based E. coli strain. *Appl. Environ. Microbiol*. **70(8)**, 4499-4504.

Finlay, B. J., & Fenchel, T. (1991). An anaerobic protozoon, with symbiotic methanogens, living in municipal landfill material. *FEMS Microbiol. Ecol.* **85**, 169–180.

Finlay, B. J. & Fenchel, T. (1996). Ecology: Role of ciliates in the natural environment. In K. Hausmann & P. C. Bradbury (Eds.), Ciliates: Cells as organisms (pp. 417–440). Stuttgart: Gustav Fischer Verlag.

Finn, R. D., Bateman, J., Clements, P., Coggill, R. Y., Eberhardt, S. R., Eddy, A., Heger, K. *et al*. (2014). The Pfam protein families database. *Nucl. Acid. Res.* **42**, 222-230.

Foissner, W. (1987). Soil protozoa: Fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Prog. Protistol.* **2**, 69–212.

Fontes, C. M. G. A. & Gilbert, H. J. (2010). Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu. Rev. Biochem*. **79**, 655-681.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W. & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 3805-3810.

Frickmann, H., Zautner, A. E., Moter, A., Kikhney, J., Hagen, R. M., Stender, H. & Poppert, S. (2017). Fluorescence in situ hybridization (FISH) in the microbiological diagnostic routine laboratory: a review. *Crit. Rev. Microbiol*. **43(3)**, 263-293.

Gabor, E. M., Alkema, W. B. & Janssen, D. B. (2004). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ. Microbiol*. **6(9)**, 879-886.

Galand, P. E., Casamayor, E. O., Kirchman, D. L. & Lovejoy, C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22427-22432.

García, C. M., & Niell, F. X. (1993). Seasonal changes in a saline temporary lake

(Fuenta de Piedra, southern Spain). *Hydrobiologia* **267** , 211–223.

Garrity, G.M. & Holt, J.G. (2001). Phylum BVI. *Chloroflexi* phy. nov.. *In* D.R. Boone & R.W. Castenholz (eds.), Vol. 1: The *Archaea* and the deeply branching and phototrophic *Bacteria*. *In* G.M. Garrity (ed.), *Bergey's Manual of Systematic Bacteriology*, 2nd ed., Springer-Verlag, New York.: 427-446.

Geng, A., Zou, G., Yan, X., Wang, Q., Zhang, J., Liu, F., Zhu, B. & Zhou, Z. (2012). Expression and characterization of a novel metagenome-derived cellulase Exo2b and its application to improve cellulase activity in Trichoderma reesei. *Appl. Microbiol. Biotechnol.* **96**, 951-962.

George, D. G., Talling, J. F. & Rigg, E. (2000). Factors influencing the temporal coherence of five lakes in the English Lake District. *Freshwater Biol.* **43**, 449-461.

Gerlach, W., Junemann, S., Tille, F., Goesmann, A. & Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* **10**, 430-439.

Giannoukos, G., Ciulla, D. M., Huang, K., Haas, B. J., Izard, J., Levin, J. Z., Livny, J. Earl, A. M. *et al*. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, r23.

Gifford, S. M., Sharma, S., Rinta-Kanto, J. M. & Moran, M. A. (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME J.* **5**, 461-472.

Gijzen, H. J., & Barugahare, M. (1992). Contribution of anaerobic protozoa and methanogens to hindgut metabolic activities of the American cockroach, *Periplaneta americana. Appl. Environ. Microbiol.* **58**, 2565–2570.

Gijzen, H. J., Drift, C., Barugahare, M. & Camp, H. J. M. (1994). Effect of host diet and hindgut microbial composition on cellulolytic activity in the hindgut of the American cockroach *Periplaneta americana*. *Appl. Environ. Microbiol.* **60 (6)**, 1822-1826.

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S. & Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environ. Microbiol*. **11(12)**, 3132-3139.

Gilbert, H. J., Know, J. P. & Boraston, A. B. (2013). Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr. Opin. Struct. Biol.* **23**, 669-677.

Glaus, M. A. & Van Loon, L. R. (2008). Degradation of cellulose under alkaline conditions: new insights from a 12 years degradation study. *Environ. Sci. Technol.* **42(8)**, 2906-2911.

Glaus, M. A., Van Loon, L. R., Schwyn, B., Vines, S., Williams, S. J., Larsson, P. & Puigdomenech, I. (2008). Long-term predictions of the concentration of a-isosaccharinic acid in cement pore water. *MRS Proceedings* **1107**, 605-612.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resources* **11**, 759-769.

Gocmen, B., Dehority, B. A., Talu, G. H., & Rastgeldy, S. (2001). The rumen ciliate fauna of domestic sheep (Ovis ammon aires) from the Turkish Republic of Northern Cyprus. *J. Eukaryot. Microbiol.* **48**, 455–459.

Godfrey, T. & West, S. eds (1996). *Industrial enzymology* (2nd edn), Macmillan Press Ltd.

Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methe, B. A. & Yooseph, S. (2010). METAREP: JCVI metagenomics reports—an open source tool for high performance comparative metagenomics. *Bioinformatics* **26(20)**, 2631-2632.

Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* **3**, 1314-1317.

Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. (2008). Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene* **407(1-2)**, 30-41.

Goodwin, S., McPherson, J. D. & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Rev. Genet.* **17**, 333-351.

Gould, J. M. (1984). Alkaline peroxide delignification of agricultural residues to enhance enzymatic saccharification. *Biotechnol. Bioeng*. **26**, 46-52.

Gould, J. M. (1985). Studies on the mechanism of alkaline peroxide delignification of agricultural residues. *Biotechnol. Bioeng*. **27**, 225-231.

Grant, W. D., Holtom, G. J., Rosevear, A. & Widdowson, D. (1997). A review of environmental microbiology relevant to the disposal of radioactive waste in a deep underground repository. Nirex Report NSS/R329.

Grant, W., Greedy, R., Holtom, G. J., O'Kelly, N., Rosevear, A. & Widdowson, D. (2001). The survival of microorganisms in a deep cementitious repository under alkaline, high temperature conditions. AEA Technology Report AEAT/R/ENV/0227.

Grant, W. D. & Sorokin, D. Y. (2011). Distribution and diversity of soda lake alkaphiles. In: Horikoshi K, Antranikian G, Bull AT, Robb FT, Stetter KO (Eds) Extremophiles Handbook, Volume 1 Springer, Tokyo, pp 27-54.

Gray, N. D., Hastings, R. C., Sheppard, S. K., Loughnane, P., Lloyd, D., McCarthy, A. J. & Head, I. M. (2003). Effects of soil improvement treatments on bacterial community structure and soil processes in an upland grassland soil. *FEMS Microbiol. Ecol.* **46**, 11-22.

Greenfield, B. F., Hurdus, M. H., Spindler, M. W. & Thomason, H. P. (1997). The effects of the products from the anaerobic degradation of cellulose on the colubility and sorption of radioelements on the near field. Nirex report NSS/R376.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. & Bailey, M. J. (2000). Rapid method for co-extraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA- based microbial community composition. *Appl. Environ. Microbiol.* **66(12)**, 5488-5491.

Guirong Su, N. R., Hua, Z. X., Zhu, S., & Imai, S. (2000). Rumen ciliated protozoan fauna of the yak (*Bos grunniens*) in China with the description of *Entodinium monuon*. sp. *J. Eukaryot. Microbiol.* **47**, 178–182.

Gurijala, K. R. & Suflita. J. M. (1993). Environmental factors influencing methanogenesis from refuse in landfill samples. *Environ. Sci. Tech.* **27**, 1176-1181.

Gusakov, A. V. (2011). Alternatives to *Trichoderma reesei* in biofuel production. *Trends Biotechnol.* **29(9)**, 419-425.

Hackstein, J. H. P., & Stumm, C. K. (1994). Methane production by terrestrial arthropods. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 5441–5445.

Haitjema, C. H., Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. (2014). Anaerobic gut fungi: Advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. *Biotechnol. Bioeng.* **111(8)**, 1471-1482.

Hamraeus, G., von Wachenfeldt, C. & Hederstedt, L. (2003). Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. *Mol. Gen. Genomics* **269**, 706-714.

Harhangi, H. R., Akhmanova, A. S., Emmens, R., van der Drift, C., de Laat, W. T., van Dijken J. P. et al. (2003). Xylose metabolism in the anaerobic fungus *Piromyces* sp. strain E2 follows the bacterial pathway. *Arch. Microbiol*. **180(2)**, 134-141.

Hasunuma, T., Okazaki, F., Okai, N., Hara, K. Y., Ishii, J. & Kondo, A. (2013). A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. *Bioresour. Technol*. **135**, 513-522.

He, S. M., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., Wang, Z., Chen, F., Lindquist, E. A., Sorek, R. & Hugenholtz, P. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* **7**, 807-858.

Heath, T. G. & Williams, S. J. (2005). Effects of organic complexants and their treatment in performance assessments. Serco Assurance Report SA/ENV-0726.

Hemsworth, G. R., Johnston, E. M., Davies, G. J. & Walton, P. H. (2015). Lytic polysaccharide monooxygenases in biomass conversion. *Trends Biotechnol*. **33(12)**, 747-761.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J*. **280**, 309-316.

Henrissat, B. & Bairoch, A. (1993). New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J*. **293**, 781-788.

Henrissat, B. & Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol*. **7**, 637-644.

Hess, M., Sczyrba, A., Egan, R., Kim, T., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S. *et al*. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331(6016)**, 463-467.

Hill, P., Heberlig, G. W. & Boddy, C. N. (2017). Sampling terrestrial environments for bacterial polyketides. *Molecules* **22(5)**, 707.

Himmel, M. E. & Picataggio, S. K. (2008). Our challenge is to acquire deeper understanding of biomass recalcitrance and conversion. In: *Biomass recalcitrance: deconstructing the plant cell wall for bioenergy* (Himmel M. E. ed.), pp 1-6, Blackwell Publishing, Oxford.

Ho, Y. W. & Barr, D. J. S. (1995). Classification of anaerobic gut fungi from herbivores with emphasis on rumen fungi from Malaysia. *Mycologia* **87(5)**, 655-677.

Hobson, P. N. & Wallace, R. J. (1982). Microbial ecology and activities in the rumen. *Crit. Rev. Microbiol.* **9**, 165-225.

Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. (2009). Orphelia: Predicting genes in metagenomic sequencing reads. *Nucleic Acids Res*. **37**, 101-105.

Holben W. E. (2011). GC fractionation allows comparative total microbial community analysis, enhances diversity assessment, and facilitates detection of minority populations of minority populations of bacteria. In: Bruijn FJ (ed) Handbook of molecular microbial ecology I: metagenomics and complementary approaches. Wiley, Hoboken, pp 183–196.

Hong, S., Bunge, J., Leslin, C., Jeon, S. & Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* **3**, 1365-1373.

Humphreys, P., Laws, A. and Dawson, J. (2010) A Review of Cellulose Degradation and the Fate of Degradation Products Under Repository Conditions Cumbria, UK: Nuclear Decommissioning Authority (NDA).

Hungate, R. E. (1942). The culture of *Eudiplodinium neglectum*, with experiments on the digestion of cellulose. *The Biological Bulletin* **83(3)**, 303-319.

Hungate, R. E. (1950). The anaerobic mesophilic cellulolytic bacteria. *Bacteriol. Rev*. **14(1)**, 1-49.

Hungate, R. E. (1966). *The rumen and its microbes*. Academic Press New York and London.

Huson, D. H., Mitra, S., Weber, N. Ruscheweyh, H. & Schuster, S. C. (2007). Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**, 1552-1560.

Huson, D. H. & Mitra, S. (2012). Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. *Evolutionary Genomics: Statistical and Computational Methods Vol 2* **856**, 415-429.

Huttenhower, C. *et al*. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214.

Ip, C. L. et al. (2015). MinION analysis and reference consortium: Phase 1 data release and analysis. *F1000Research* **4**, 1075.

Ishii, K. & Fukui, M. (2001). Optimisation of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Appl. Environ. Microbiol*. **67(8)**, 3753-3755.

Ito, A., & Imai, S. (2000). Ciliates from the cecum of capybara (*Hydrocheorus hydrochaeris*) in Bolivia. 1. The families Hydrochoerellidae n. fam., Protohalliidae, and Pcynotrichidae. *Eur. J. Protistol.* **36**, 53–84.

Jacquet, S., Havskum, H., Thingstad, T. F. & Vaulot, D. (2002). Effects of inorganic and organic nutrient addition on a coastal microbial community (Isefjord, Denmark). *Mar. Ecol. Prog. Ser.* **228**, 3-14.

Jalak, J., Kurasin, M., Teugjas, H. & Valjamae, P. (2012). Endo-exo synergism in cellulose hydrolysis revisited. *J. Biol. Chem.* **287 (34)**, 28802-28815.

Jaramillo, P. & Matthews, H. S. (2005). Landfill-gas-to-energy projects: Analysis of net private and social benefits. *Environ. Sci. Tech.* **39**, 7365-7373.

Jenkinson, D. S., Adams, D. E. & Wild, A. (1991). Model estimates of CO2 emissions from soil in response to global warming. *Nature* **351**, 304-306.

Jeon, J. H., Kim, J. T., Kang, S. G., Lee, J. H. & Kim, S. J. (2009). Characterization and its potential application of two esterases derived from the arctic. *Mar. Biotechnol*. **11(3)**, 307-316.

Joblin, K. N. (1981). Isolation, enumeration, and maintenance of rumen anaerobic fungi in roll tubes. *Appl. Environ. Microbiol*. **42(6)**, 1119-1122.

Johansen, K. S. (2016). Discovery and industrial applications of lytic polysaccharide mono-oxygenases. *Biochem. Soc. Trans*. **44(1)**, 143-149.

Jouany, J. P. (1994). Manipulation of microbial activity in the rumen. *Arch. Anim. Nutr.* **46**,133–153.

Kakirde, K. S., Parsley, L. C. & Liles, M. R. (2010). Size does matter: Application driven approaches for soil metagenomics. *Soil Biol. Biochem*. **42(11)**, 1911-1923.

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28(1)**, 27-30.

Kang, D. D., Froula, J., Egan, R. & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *Peer J.* **3**, e1165.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S. *et al*. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28(12)**, 1647-1649.

Kemp, P., Lander, D. J. & Orpin, C. G. (1984). The lipids of the rumen fungus *Piromonas communis*. *J. Gen. Microbiol.* **130**, 27-37.

Kent, W. J. (2002). BLAT- the BLAST-like alignment tool. *Genome Res*. **12(4)**, 656-664.

Kim, D., Kim, S. N., Baik, K. S., Park, S. C., Lim, C. H., Kim, J. O. *et al*. (2011). Screening and characterization of a cellulase gene from the gut microflora of abalone using metagenomic library. *J. Microbiol.* **49(1)**, 141-145.

Kjeldsen, P., Barlaz, M. A., Rooker, A. P., Baun, A., Ledin. A. & Christensen, T. H. (2002). Present and long-term composition of MSW landfill leachate: A review. *Crit. Rev. Environ. Sci. Technol.* **32**, 297-336.

Knill, C. J. & Kennedy, J. F. (2003). Degradation of cellulose under alkaline conditions. *Carbohydrate Polymers* **51(3)**, 281-300.

Kodzius, R. & Gojobori, T. (2016). Single-cell technologies in environmental omics. *Gene* **576(2)**, 701-707.

Koetschan, C., Kittelmann, S., Lu, J., Al-Halbouni, D., Jarvis, G. N., Muller, T., Wolf, M. & Janssen, P. H. (2014). Internal transcribed spacer 1 secondary strycture analysis reveals a common core throughout the anaerobic fungi (Neocallimastigomycota). *PLoS ONE* **9(3)**, e91928.

Kopylova, E., Noe, L. & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28(24)**, 3211-3217.

Krause, D. O., Denman, S. E., Mackie, R. I., Morrison, M., Rae, A. L., Attwood, G. T. & McSweeney, C. S. (2003). Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. *FEMS Microbiol. Rev.* **27**, 663-693.

Krulwich, T. A. (1995). Alkaliphiles: 'basic' molecular problems of pH tolerance and bioenergetics. *Mol. Microbiol*. **15(3)**, 403-410.

Kubicek, C. P. *et al*. (2009). Metabolic engineering strategies for the improvement of cellulase production by *Hypocrea jecorina*. *Biotechnol. Biofuels* **2**, 19.

Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. (2009). Coding sequence determinants of gene expression in *E. coli. Science* **324(5924)**, 255-258.

Kudo, H., Cheng, K. J., & Costerton, J. W. (1987). Interactions between Treponema bryantii and cellulolytic bacteria in the in vitro degradation of straw cellulose. *Can. J. Microbiol*. **33(3),** 244-248.

Kuhad, R. C., Gupta, R. & Singh, A. (2011). Microbial cellulases and their industrial applications. *Enzyme Res.* **2011**.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12**, 118-123.

Kurakov, A. V., Khidirov, K. S., Sadykova, V. S. & Zvyagintsev, D. G. (2011). Anaerobic growth ability and alcohol fermentation activity of microscopic fungi. *Appl. Biochem. Microbiol.* **47 (2)**, 169-175.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10(3)**, R25.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359.

Laserson, J., Jojic, V. & Koller, D. (2011). Genovo: *De novo* assembly for metagenomes. *J. Comput. Biol*. **18(3)**, 429-443.

Laureano-Perez, L., Teymouri, F., Alizadeh, H. & Dale, B. E. (2005). Understanding factors that limit enzymatic hydrolysis of biomass-characterisation of pretreated corn stover. *Appl. Biochem. Biotechnol*. **121-124**, 1081-1099.

Leggio, L. L., Simmons, T. J., Poulsen, J. C., Frandsen, K. E., Hemsworth, G. R., Stringer, M. A., von Freiesleben, P., Tovborg, M. *et al*. (2015). Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. *Nat. Commun*. **6**, 5961.

Lehembre, F., Doillon, D., David, E., Perrotto, S., Baude, J., Foulon, J., Harfouche, L., Vallon, L., Poulain, J., Da Silva, C., Wincker, P., Oger-Desfeux, C., Richaud, P., Colpaert, J. V., Chalot, M., Fraissinet-Tachet, L., Blaudez, D. and Marmeisse, R. (2013). Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes. *Environ. Microbiol.* doi: 10.1111/1462-2920.12143.

Leschine, S. B. (1995). Cellulose degradation in anaerobic environments. *Annu. Rev. Microbiol*. **49**, 399-426.

Lesniewski, R. A., Jain, S., Anantharaman, K., Schloss, P. D. & Dick, G. J. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J*. **6(12)**, 2257-2268.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels* **6(1)**, 41.

Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y, Mao Y (2004). Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. *J. Biotechnol.* **107**(1),19–28.

Li, S., Xu, L. H., Hua, H., Ren, C. A. & Lin, Z. L. (2007). A set of UV-inducible autolytic vectors for high throughput screening. *J. Biotechnol*. **127(4)**, 647-652.

Li, M., Wang, B. H., Zhang, M. H., Rantalainen, M., Wang, S. Y., Zhou, H. K., Zhang, Y. et al. (2008). Symbiotic gut microbes modulate human phenotypes. *Proc. Natl. Acad. Sci. U. S. A*. **105(6)**, 2117-2122.

Li, H. & Durbin, R. (2009). Fast and accurate short rad alignment with Burrows-Wheeler-Transform. *Bioinformatics* **25**, 1754-1760.

Li, L. L., McCorkle, S. R., Monchy, S., Taghavi, S. & van der Lelie, D. (2009). Bioprospecting metagenomes: Glycosyl hydrolases for converting biomass. *Biotechnology for Biofuels* **2,** 10-20.

Liang, J., Cai, W. & Sun, Z. (2014). Single-cell sequencing technologies: Current and future. *J. Genet. Genomics* **41(10)**, 513-528.

Liggenstoffer, A. S., Youssef, N. H., Couger, M. B. & Elshahed, M. S. (2010). Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. *ISME J.* **4(10)**, 1225-1235.

Ljungdahl, L. G. & Eriksson, K. E. (1985). Ecology of microbial cellulose degradation. *Adv. Microb. Ecol*. **8**, 237-299.

Ljungdahl, L. G. (2008). The cellulase/hemicellulase system of the anaerobic fungus *Orpinomyces* PC-2 and aspects of its applied use. *Ann. NY Acad. Sci.* **1125(1)**, 308-321.

Lockhart, R. J., Van Dyke, M. I., Beadle, I. R., Humphreys, P. & McCarthy, A. J. (2006). Molecular biological detection of anaerobic gut fungi (*Neocallimastigales*) from landfill sites. *Appl. Environ. Microbiol*. **72(8)**, 5659-5661.

Logares, R., Haverkamp, T. H. A., Kumar, S., Lanzen, A., Nederbragt, A. J., Quince, C. & Kauserud, H. (2012). Environmental microbiology through the lens of high-

throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *J. Microbiol. Methods* **91**, 106-113.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnol.* **30**, 434-439.

Lombard, V. *et al*. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J*. **432**, 437-444.

Lowe, S. E., Theodorou, M. K., Trinci, A. P. J. & Hespell, R. B. (1985). Growth of anaerobic rumen fungi on defined and semi-defined media lacking rumen fluid. *J. Gen. Microbiol.* **131**, 2225-2229.

Lowe, S. E., Theodorou, M. K. & Trinci, A. P. J. (1987). Growth and fermentation of an anaerobic ruminal fungus on various carbon sources and the effect of temperature on development. *Appl. Environ. Microbiol.* **53**, 1210-1215.

Luo, C. W., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *Plos One* **7(2)**, e30087.

Luther, R., Trenkle, A. & Burroughs, W. (1966). Influence of rumen protozoa on volatile acid production and ration digestibility in lambs. *J. Anim. Sci.* **25(4)**, 1116-1122.

Lynch, M. D. & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol*. **13(4)**, 217-229.

Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. (2002). Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506.

Lynd, L. R., van Zyl, W. H., McBride, J. E. & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: An update. *Curr. Opin. Biotechnol.* **16**, 577-583.

Lynn, D. H. (1996a). My journey in ciliate systematics. *J. Euk. Microbiol.* **43**, 253–260.

Lynn, D. H. (2010). Subphylum 2. Intramacronucleata: Class 2. Armophorea. In *The ciliated Protozoa. Characterisation, classification and guide to the literature*. (pp 176-179). Pergamon press, New York.

Lynn, D. H., & Corliss, J. O. (1991). Ciliophora. In F. W. Harrison & J. O. Corliss (Eds.), Microscopic anatomy of invertebrates (Vol. 1, chap. 5, pp. 333–467). New York: Wiley−Liss.

Lynn, D. H., & Montagnes, D. J. S. (1991). Global production of heterotrophic marine planktonic ciliates. In P. C. Reid, C. M. Turley, & P. H. Burkhill (Eds.), Protozoa and their role in marine processes, NATO Publications (Vol. G25, pp. 281–307). Berlin: Springer.

Lynn, D. H., & Small, E. B. (1981). Protist kinetids: Structural conservatism, kinetid structure, and ancestral states. *Biosystems* **14**, 317-322.

Lynn, D. H. & Wright, A. G. (2013). Biodiversity and molecular phylogeny of Australian *Clevelandella* species, intestinal endosymbiotic ciliates in the wood-feeding roach *Panesthis cribrata* Saussure, 1864. *J. Euk. Microbiol.* **60**, 335-341.

Mader, U., Nicolas, P., Richard, H., Bessieres, P. & Aymerich, S. (2011). Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Curr. Opin. Biotechnol.* **22**, 32-41.

Marchessault, R. H. & Sundararajan, P. R. (1983). Cellulose. In *The polysaccharides*. (pp 11-95). New York: Academic.

Mardis, E. R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols* **12**, 213-218.

Martin-Cuadrado, A. B., Lopez-Garcia, P., Alba, J. C., Moreira, D., Monticelli, L., Strittmatter, A. *et al*. (2007). Metagenomic of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One* **2(9)**, e914.

Martinez, A., Bradley, A. S., Waldbauer, J. R., Summons, R. E. & DeLong, E. F. (2007). Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc. Natl. Acad. Sci. U. S. A.* **104(13)**, 5590-5595.

Martinez, A., Tyson, G. W. & DeLong, E. F. (2010). Widespread known and novel phosphonate utilisation pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* **12 (1)**, 222-238.

Markovitz, V. M., Chen, I. M., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I. et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res*. **42**, D568-573.

Marx, V. (2013). PCR: living life amplified and standardized. *Nature Methods* **10(5)**, 391-395.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics* 13, **31**. Available: http://www.biomedcentral.com/1471-2105/13/31

Mason, U. O. *et al*. (2012). Metagenome, metatranscriptome and single cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* **6**, 1715-1727.

McCarren, J., Becker, J. W., Repeta, D. J., Shi, Y., Young, C. R., Malmstrom, R. R., Chisholm, S. W. & DeLong, E. F. (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *PNAS* **107 (38)**, 16420-16427.

McDonald, J. E., Lockhart, R. J., Cox, M. J., Allison, H. E. & McCarthy, A. J. (2008). Detection of novel Fibrobacter populations in landfill sites and determination of their relative abundance via quantitative PCR. *Environ. Microbiol.* **10**, 1310-1319.

McDonald, J. E., de Menezes, A. B., Allison, H. E. & McCarthy, A. J. (2009). Molecular biological detection and quantification of novel Fibrobacter populations in freshwater lakes. *Appl. Environ. Microbiol.* **75**, 5148-5152.

McDonald, J. E., Houghton, J. N. I., Rooks, D. J., Allison, H. E. & McCarthy, A. J. (2012). The microbial ecology of anaerobic cellulose degradation in municipal waste landfill sites: evidence for a role of Fibrobacters. *Environ. Microbiol.* **152**, 2003-2012.

Mendes, R., Garbeva, P. & Raaijmakers, J. M. (2013). The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol. Rev*. **37**, 634-663.

Merino, S. T. & Cherry, J. (2007). Progress and challenges in enzyme development for biomass utilization. *Adv. Biochem. Eng. Biotechnol*. **108**, 95-120.

Mettel, C., Kim, Y., Shrestha, P. M. & Liesack, W. (2010). Extraction of mRNA from Soil. *Appl. Environ. Microbiol.* **76**, 5995-6000.

Mewis, K., Taupp, M., & Hallam, S. J. (2011). A high throughput screen for biomining cellulase activity from metagenomic libraries. *Journal of visualized experiments: JoVE*, (**48)**.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodrigues, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R. A. (2008). The metagenomic RAST server- A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386-393.

Michalet-Doreau, B., Fernandez, I., Peyron, C., Millet, L. & Fonty, G. (2001). Fibrolytic activities and cellulolytic bacterial community structure in the solid and liquid phases of rumen contents. *Reprod. Nutr. Dev.* **41**, 187-194.

Michalowski, T., Belzecki, G., Kwiatkowska, E., & Pajak, J. J. (2003). The effect of selected rumen fauna on fibrolytic enzyme activities, bacterial mass, fibre disappearance and fermentation pattern in sheep. *J. Anim. Feed Sci.* **12**, 45–64.

Montella, S., Ventorino, V., Lombard, V., Henrissat, B., Pepe, O. & Faraco, V. (2017). Discovery of genes coding for carbohydrate-active enzyme by metagenomic analysis of lignocellulosic biomasses. *Sci. Rep*. 7:42623.

Mooney, C. A., Mansfield, S. D., Touhy, M. G. & Saddler, J. N. (1998). The effect of initial pore volume and lignin content on the enzymatic hydrolysis of softwoods. *Bioresource Technol*. **64**, 113-119.

Moran, M. A. (2009). Metatranscriptomics: Eavesdropping on complex microbial communities. *Microbe* **4**, 329-335.

Morrison, M., Pope, P. B., Denman, S. E. & McSweeney, C. S. (2009). Plant biomass degradation by gut microbiomes: more of the same or something new? *Curr. Opin. Biotechnol.* 20, 358-363.

Moseir, N., Hendrickson, R., Ho, N., Sedlak, M. & Ladisch, M. R. (2005). Optimisation of pH controlled liquid hot water pretreatment of corn stover. *Bioresour. Technol*. **96**, 1986-1993.

Munir, R. & Levin, D. B. (2016). Enzyme systems of anerobes for biomass conversion. *Adv. Biochem. Eng. Biotechnol*. **156**, 113-138.

Nacke, H., Engelhaupt, M., Brady, S., Fischer, C., Tautzt, J. & Daniel, R. (2012). Identification and characterisation of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnol. Lett.* **34 (4)**, 663-675.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. (2012). MetaVelvet: An extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucl. Acids Res*. **40(20)**, 1-12.

Narayanan, N., Priya, M., Haridas, A. & Manilal, V. B. (2007). Isolation and culturing of a most common anaerobic ciliate, Metopus sp. *Anaerobe* **13(1)**, 14-20.

Neufeld, J. D., Schafer, H., Cox, M. J., Boden, R., McDonald, I. R. & Murrell, J. C. (2007). Stable-isotope probing implicates *Methylophaga* spp and novel *Gammaproteobacteria* in marine methanol and methylamine metabolism. *ISME J.* **1**, 480-491.

Nguyen, N. H., Maruset, L., Uengwatwanit, T., Mhuantong, W., Harnpicharnchai, P., Champreda, V *et al*. (2012). Identification and characterisation of a cellulase-encoding gene from the buffalo rumen metagenomic library. *Biosci. Biotechnol. Biochem.* **76(6)**, 1075-1084.

Ni, Y., Li, J. & Panagiotou, G. (2016). COMAN: A webserver for comprehensive metatranscriptomics analysis. *BMC Genomics* **17**, 622.

Nicholson, M. J., Theodorou, M. K. & Brookman, J. L. (2005). Molecular analysis of the anaerobic rumen fungus *Orpinomyces*- Insights into an AT-rich genome. *Microbiology* **151(1)**, 121-133.

Nicol, G. W., Campbell, C. D., Chapman, S. J. & Prosser, J. I. (2007). Afforestation of moorland leads to changes in crenarchaeal community structure. *FEMS Microbiol. Ecol.* **60**, 51-59.

Nieves, R. A. *et al*. (1998). Technical communication: Survey and analysis of commercial cellulase preparations suitable for biomass conversion to ethanol. *World J. Microbiol. Biotechnol.* **14**, 301-304.

Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K. H. (2008). Intraspecific ITS variability in the kingdom fungi as expressed in the international

sequence databases and its implications for molecular species identification. *Evol. Bioinform. Online* **4**, 193-201.

Nimchua, T., Thongaram, T., Uengwetwanit, T., Pongpattanakitshote, S. & Eurwilaichitr, L. (2012). Metagenomic analysis of novel lignocellulose-degradaing enzymes from the higher termite guts inhabiting microbes. *J. Microbiol. Biotechnol.* **22(4)**, 462-469.

Nirex, *Generic Disposal System Specification Volume 1 - Specification*, Generic Repository Studies, Nirex Report N/075, 2003.

Nirex, *The Viability of a Phased Geological Repository Concept for the Long-term Management of the UK's Radioactive Waste*, Nirex Report N/122, 2005.

O'Malley, M. A., Theodorou, M. K. & Kaiser, C. A. (2012). Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes native to *Piromyces* sp E2 in *Saccharomyces cerevisiae*. *Environ. Prog. Sustain Energy* **31(1)**, 37-46.

O'Sullivan, A. C. (1997). Cellulose: the structure slowly unravels. *Cellulose* **4**, 173-207.

Ohkuma, M., & Kudo, T. (1996). Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. *Appl. Environ. Microbiol.* **62(2),** 461-468.

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. (1986). Microbial ecology and evolution- A ribosomal-RNA approach. *Ann. Rev. Microbiol.* **40**, 337-365.

Olson, D. G., McBride, J. E., Shaw, J. & Lynd, L. R. (2012). Recent progress in consolidates bioprocessing. *Curr. Opin. Biotechnol*. **23**, 396-405.

Orpin, C. G. (1975). Studies on the rumen flagellate *Neocallimastix frontalis*. *J. Gen. Microbiol.* **91(2)**, 249-262.

Ottesen, E. A., Marin, R, Preston, C. M., Young, C. R., Ryan, J. P., Scholin, C. A. & DeLong, E. F. (2011). Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J.* **5 (12)**, 1881-1895.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691-5702.

Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal-RNA sequences. *Adv. Microb. Ecol.* **9**, 1-55.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276(5313)**, 734-740.

Palackal, N., Lyon, C. S., Zaidi, S., Luginbuhl, P., Dupree, P., Goubet, F., Macomber, J. L. *et al*. (2007). A multifunctional hybrid glycosyl hydrolase discovered in an uncultured microbial consortium from ruminant gut. *Appl. Microbiol. Biotechnol.* **74(1)**, 113-124.

Palmisano, A. C. & Barlaz, M. A. (1996). Introduction to solid waste decomposition. In *Microbiology of solid waste*. Palmisano, A. C., Barlaz, M. A. (eds): CRC Press.

Parsley, L. C., Consuegra, E. J., Kakirde, K. S., Land, A. M., Harper, W. F. & Liles, M. R. (2010). Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl. Environ. Microbiol*. **76(11)**, 3753-3757.

Patzek, T. W. & Pimentel, D. (2005). Thermodynamics of energy production from biomass. *Crit. Rev. Plant Sci.* 24, 327-364.

Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ Jr, Davies PF (2003). Fidelity of enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. *Physiol. Genomics* **13**, 147–156.

Polz, M. F. & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64**, 3724-3730.

Poretsky, R. S., Gifford, S., Rinta-Kanto, J., Vila-Costa, M. & Moran, M. A. (2009). Analysing gene expression from marine microbial communities using environmental transcriptomics. JoVE. 24. http://www.jove.com/index/Details.stp?ID=1086, doi: 10.3791/1086.

Prade, R. A. (1996). Xylanases: From biology to biotechnology. *Biotechnol. Genet. Eng. Rev.* **13(1)**, 101-132.

Prakash, T. & Taylor, T. D. (2012). Functional assignment of metagenomic data: Challenges and applications. *Briefings in Bioinformatics* **13**, 711-727.

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R. & Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell* **160(6)**, 1111-1124.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. & Glockner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35(21)**, 7188-7196.

Quinlan, R. J., Sweeney, M. D., Leggio, L. L., Otten, H., Poulsen, J. N., Johansen, K. S., Krogh, K. B. R. M. et al. (2011). Insights into the oxidative degradation of cellulose

by a copper metalloenzyme that exploits biomass components. *PNAS* **108(37)**, 15079-15084.

Radax, R., Rattei, T., Lanzen, A., Bayer, C., Rapp, H. T., Urich, T. & Schleper, C. (2012). Metatranscriptomics of the marine sponge Geodia barretti: Tackling phylogeny and function of its microbial community. *Environ. Microbiol.* **14**, 1308-1324.

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cariney, J., Eckert, C. A. *et al*. (2006). The path forward for biofuels and biomaterials. *Science* **311**, 484-489.

Ransom-Jones, E., Jones, D. L., McCarthy, A. J. & McDonald, J. E. (2012). The *Fibrobacteres*: An important phylum of cellulose-degrading bacteria. *Microbial. Ecol.* **63**, 267-281.

Rashamuse, K. J., Visser, D. F., Hennessy, F., Kemp, J., Roux-van der Merwe, M. P., Badenhorst, J., Ronneburg, T., Francis-Pope, R. & Brady, D. (2013). Characterisation of two bifunctional cellulase-xylanase enzymes isolated from a bovine rumen metagenome library. *Curr. Microbiol.* **66**, 145-151.

Redon, E., Loubiere, P. & Cocaign-Bousquet, M. (2005). Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *J. Biol. Chem.* **280**, 36380-36385.

Renewable Fuels Association (2007). Ethanol industry outlook 2007: building new horizons. *Renewable Fuels Administration*; www.ethanolrfa.org/resource/outlook

Renou, S., Givaudan, J. G., Poulain, S., Dirassouyan, F. & Moulin, P. (2008). Landfill leachate treatment: Review and opportunity. *J. Hazardous Materials* **150**, 468-493.

Rhee, J. K., Ahn, D. G., Kim, Y. G. & Oh, J. W. (2005). New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Appl. Environ. Microbiol.* **71(2)**, 817-825.

Rho, M., Tang, H. & Ye, Yuzhen. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res*. **38(20)**, e191.

Rieder, S. R. & Frey, B. (2013). Methyl-mercury affects microbial activity and biomass, bacterial community structure but rarely the fungal community structure. *Soil Biol. Biochem.* **64**, 164-173.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P. & Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437.

Roberts, J. D., Preston, B. D., Johnston, L. A., Soni, A., Loeb, L. A. & Kunkel, T. A. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol. Cell. Biol*. **9(2)**, 469-476.

Rooks, D. J., McDonald, J. E. & McCarthy, A. J. (2012). Metagenomic approaches to the discovery of cellulases. *Methods Enzymol*. **510**, 375-394.

Rout SP, Radford J, Laws AP, Sweeney F, Elmekawy A, Gillie LJ, et al. (2014). Biodegradation of the alkaline cellulose degradation products generated during radioactive waste disposal. *PLoS ONE* **9**.

Rout, S. P., Charles, C. J., Garratt, E. J., Laws, A. P., Gunn, J. & Humphreys, P. N. (2015). Evidence of the generation of isosaccharinic acids and their subsequent degradation by local microbial consortia within hyperalkaline contaminated soils, with relevance to intermediate level radioactive waste disposal. *PLoS ONE* **10(3)**, e0119164.

Rubin, E. M. (2008). Genomics of cellulosic biofuels. *Nature* **454(14)**, 841-845.

Russell, J. B., Muck, R. E. & Weimer, P. J. (2009). Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen. *FEMS Micobiol. Ecol*. **67**, 183-197.

Sajith, S., Priji, P., Sreedevi, S. & Benjamin, S. (2016). An overview on fungal cellulases with an industrial perspective. *J. Nutr. Food Sci*. **6**, 1-13.

Sanderson, K. (2006). US biofuels: A field in ferment. *Nature* **444**, 673-676.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A*. **74**, 5463-5467.

Schipper, C., Hornung, C., Bijtenhoorn, P., Quitschau, M., Grond, S. & Streit, W. (2009). Metagenome-derived clones encoding two novel lactonase family proteins involved in biofilm inhibition in *Pseudomonas aeruginosa. Appl. Environ. Microbiol*. **75(1)**, 224-233.

Schloss, P., D. & Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* **14(3)**, 303-310.

Schmieder, R. & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864.

Scholz, M. B., Lo, C. & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotech.* **23**, 9-15.

Schubert, C. (2006). Can biofuels finally take centre stage? *Nat. Biotechnol*. **24**, 777-784.

Schurz, J. (1978). Bioconversion of cellulosic substances into energy chemical and microbial protein symp. In: *Proc*, pp 37.

Schwarz, W. H. (2001). The cellulosome and cellulose degradation by anaerobic bacteria. *Appl. Microbiol. Biotechnol.* **56**, 634-649.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**, 811-814.

Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. & Rosenow, C. (2003). Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216-223.

Sharma, V. K., Kumar, N., Prakash, T. & Taylor, T. D. (2012). Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PLoS One* **7(4)**, e34030.

Shi, Y., Tyson, G. W. & DeLong, E. F. (2009). Metatranscriptomics reveals unique small rRNAs in the ocean's water column. *Nature* **459**, 266-269.

Shi, Y., Tyson, G. W., Eppley, J. M. & DeLong, E. F. (2011). Integrated metatranscriptomic and metagenomic analysis of stratified microbial assemblages in the open ocean. *ISME J.* **5**, 999-1013.

Shin, S. G., Lee, C. S., Hwang, K., Ahn, J. H. & Hwang, S. (2008). Use of order-specific primers to investigate the methanogenic diversity in acetate enrichment system. *J. Ind. Microbiol. Biotechnol.* **35**, 1345-1352.

Shinkai, T. & Kobayashi, Y. (2007). Localization of ruminal cellulolytic bacteria on plant fibrous materials as determined by fluorescence in situ hybridization and real-time PCR. *Appl. Environ. Microbiol.* **73**, 1646-1652.

Shuai, L., Yang, Q., Zhu, J. Y., Lu, F. C., Weimer, P. J., Ralph, J. & Pan, X. J. (2010). Comparative study of SPORL and dilute-acid pretreatments of spruce for cellulosic ethanol production. *Bioresour. Technol*. **101**, 3106-3114.

Simon, C., Herath, J., Rockstroh, S. & Daniel, R. (2009). Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl. Environ. Microbiol.* **75(9)**, 2964-2968.

Singh, B. K., & Macdonald, C. A. (2010). Drug discovery from uncultivable microorganisms. *Drug Discovery Today*, **15(17),** 792-799.

Slaytor, M. (1992). Cellulose digestion in termites and cockroaches- what role do symbionts play? *Comp. Biochem. Physiol.* **103**, 775-784.

Smith, C. J., Nedwell, D. B., Dong, L. F. & Osborn, A. M. (2006). Evaluation of quantitative polymerase chain reaction-based approaches for determining gene copy and gene transcript numbers in environmental samples. *Environ. Microbiol.* **8**, 804-815.

Smith, C. J. & Osborn, A. M. (2009). Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol. Ecol.* **67**, 6-20.

Smith, S. P. & Bayer, E. A. (2013). Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. *Curr. Opin. Struct. Biol.* **23**, 686-694.

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12115-12120.

Song, Y. H., Lee, K. T., Baek, J. Y., Kim, M. J., Kwon, M. R., Kim, Y. J., Park, M. R., Ko, H., Lee, J. S. & Kim, K. S. (2017). Isolation and characterization of a novel glycosyl hydrolase family 74 (GH74) cellulase from the black goat rumen metagenomic library. *Folia. Microbiol. (Praha)* **62(3)**, 175-181.

Sorensen, M. A., Kurland, C. G. & Pedersen, S. (1989). Codon usage determines translation rate in *E. coli. J. Mol. Biol*. **207(2)**, 365-377.

Sorokin, D. Y. (2005). Is there a limit for high-pH life? *Int. J. Syst. Evol. Microbiol*. **55(4)**, 1405-1406.

Stamps, B. W., Lyles, C. N., Siflita, J. M., Masoner, J. R., Cozzarelli, I., M., Kolpin, D. W. & Stevenson, B. S. (2016). Municipal solid waste landfills harbor distinct microbiomes. *Front. Microbiol*. **7**, 534.

Stan, K., Belzecki, G., Kasperowicz, A., Kwiatkowska, E., & Michalowski, T. (2006). The ability of the rumen ciliate Anoplodinium denticulatum to utilize hemicellulosic material for in vitro growth. *J. Anim. Feed Sci.* **15**, 39–42.

Stanton, T. B., & Canale-Parola, E. (1980). *Treponema bryantii* sp. nov., a rumen spirochete that interacts with cellulolytic bacteria. *Arch. Microbiol.* **127(2),** 145-156.

Steenbakkers, P. J. M., Li, X. L., Ximenes, E. A., Arts, J. G., Chen, H. Z., Ljungdahl, L. G. & Den Camp, H. (2001). Noncatalytic docking domains of cellulosomes of anaerobic fungi. *J. Bacteriol.* **183**, 5325-5333.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207-213.

Stevenson, D. M. & Weimer, P. J. (2007). Dominance of *Prevotella* and low abundance of classical ruminal bacterial species in the bovine rumen revealed by relative quantification real-time PCR. *Appl. Microbiol. Biotechnol.* **75(1)**, 165-174.

Stewart, F. J., Ulloa, O. & Delong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* **14**, 23-40.

Stewart, F. J., Ottesen, E. A. & DeLong, E. F. (2010). Development and quantitative analysis of a universal rRNA subtraction protocol for microbial metatranscriptomics. *ISME J.* **4**, 896-907.

Suen, G., Weimer, P. J., Stevenson, D. M., Aylward, F. O., Boyum, J., Deneke, J., Drinkwater, C. *et al*. (2011). The complete genome sequence of *Fibrobacter succinogenes* S85 reveals a cellulolytic and metabolic specialist. *PLOS One* 6, e18814.

Suzuki, M. T. & Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **66**, 4605-4614.

Suzuki, M. T., Taylor, L. T. & DeLong, E. F. (2000). Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5'-nuclease assays. *Appl. Environ. Microbiol.* **66**, 4605-4614.

Tajima, K., Aminov, R. I., Nagamine, T., Ogata, K., Nakamura, M., Matsui, H. & Benno, Y. (1999). Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. *FEMS Microbiol. Ecol.* **29**, 159-169.

Tajima, K., Arai, S., Ogata, K., Nagamine, T., Matsui, H., Nakamura, M. et al. (2000). Rumen bacterial community transition during adaptation to high-grain diet. *Anaerobe* **6**, 273-284.

Tajima, K., Aminov, R. I., Nagamine, T., Matsui, H., Nakamura, M. & Benno, Y. (2001). Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Appl. Environ. Microbiol.* **67**, 2766-2774.

Takasaki, K., Miura, T., Kanno, M., Tamaki, H., Hanada, S., Kamagata, Y. & Kimura, N. (2013). Discovery of glycoside hydrolase enzymes in an Avicel adapted forest soil fungal community by a metatranscriptomic approach. *Plos One* **8(2)**, e55485.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R. *et al*. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Taupp, M., Mewis, K., & Hallam, S. J. (2011). The art and design of functional metagenomic screens. *Curr. Opin. Biotechnol.* **22(3),** 465-472.

Taylor, W. D., & Heynen, M. L. (1987). Seasonal and vertical distribution of Ciliophora in Lake Ontario. *Canadian Journal of Fisheries and Aquatic Sciences* **44**, 2185–2191.

Teather, R. M. & Wood, P.J. (1982). Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria form the bovine rumen. *Appl. Environ. Microbiol*. **43(4)**, 777-780.

Teymouri, F., Laureano-Perez, L., Alizadeh, H. & Dale, B. E. (2005). Optimisation of the ammonia fiber explosion (AFEX) treatment parameters for enzymatic hydrolysis of corn stover. *Bioresour. Technol*. **96**, 2014-2018.

Theodorou, M. K., Gill, M., Kingspooner, C., & Beever, D. E. (1990). Enumeration of anaerobic *Chytridiomycetes* as thallus-forming units- novel method for quantification of fibrolytic fungal populations from the digestive-tract ecosystem. *Appl. Environ. Microbiol.* **56**, 1073-1078.

Theodorou, M. K., Mennim, G., Davies, D. R., Zhu, W. Y., Trinci, A. P. J. & Brookman, J. L. (1996). Anaerobic fungi in the digestive tract of mammalian herbivores and their potential for exploitation. *Proc. Nutr. Soc*. **55(3)**, 913-926.

Theodorou, M. K., Brookman, J. & Trinci, A. P. J. (2005). Anaerobic fungi. *In Methods in gut microbial ecology for ruminants* (Makkar, H. P. S., McSweeney, C. S. eds.) pp 55-66. Dordrecht, Springer Netherlands.

Thomas, T., Gilbert, J. & Meyer, F. (2012). Metagenomics - A guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2**, 3-14.

Timoshenko, O., & Imai, S. (1997). Three new intestinal protozoan species of the genus *Latteuria* n. g. (Ciliophora: Trichostomatida) from Asian and African elephants. *Parasitol. Int.* **46**, 297–303.

Torsvik, V. & Ovreas, L. (2002). Microbial diversity and function in soil: From genes to ecosystems. *Curr. Opin. Microbiol*. **5(3)**, 240-245.

Trinci, A. P. J., Davies, D. R., Gull, K., Lawrence, M. I., Nielsen, B. B., Rickers, A. & Theodorou, M. K. (1994). Anaerobic fungi in herbivorous animals. *Mycological Res.* **98**, 129-152.

Tripp, H. J., Hewson, I., Boyarsky, S., Stuart, J. M. & Zehr, J. P. (2011). Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.* **39(20),** 8792-8802.

Tsai, K. P. & Calza, R. E. (1992). Enzyme-based DNA extraction from zoospores of ruminal fungi. *Fungal Genet. Newsl.* **39**, 86-88.

Tveit, A., Schwacke, R., Svenning, M. M. & Urich, T. (2013). Organic carbon transformations in high-Arctic peat soils: Key functions and microorganisms. *ISME J.* **7**, 299-311.

Uchiyama, T. & Miyazaki, K. (2009). Functional metagenomics for the enzyme discovery: Challenges to efficient screening. *Curr. Opin. Biotechnol.* **20**, 616-622.

Urich, T., Lanzen, A., Qi, J., Huson, D. H., Schleper, C. & Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the metatranscriptome. *PLoS One* **3(6)**, e2527.

US Department of Agriculture (2016). 2015 Energy balance for the corn-ethanol industry. *US Department of Agriculture*.

US Department of Energy (2015). Energy efficiency and renewable energy. *US Department of Energy.*

US Energy Information Administration (2014). Monthly ethanol production report February 2014. *US Energy Information Administration*.

Vaaje-Kolstad, G., Westereng, B., Horn, S. J., Liu, Z., Zhai, H., Sorlie, M. & Eijsink, V. G. (2010). An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. *Science* **330(6001)**, 219-222.

Van Dyke, M. I. & McCarthy, A. J. (2002). Microbial biological detection and characterisation of *Clostridium* populations in municipal landfill sites. *Appl. Environ. Microbiol*. **68(4)**, 2049-2053.

Veira, D. M., Ivan, M., & Jui, P. Y. (1983). Rumen ciliate protozoa: Effects on digestion in the stomach of sheep. *J. Dairy Sci.* **66(5)**, 1015-1022.

Veldkamp, H. (1960). Isolation and characteristics of *Treponema zuelzerae* nov. spec., an anaerobic, free-living spirochete. *Antonie van Leeuwenhoek*, ***26*(1),** 103-125.

Vogels, G. D. (1979). Global cycle of methane. *Antonie Van Leeuwenhoek J. Microbiol.* **45**, 347-352.

Voget, S., Steele, H. L. & Streit, W. R. (2006). Characterization of a metagenome-derived halotolerant cellulase. *J. Biotechnol.* **126(1)**, 26-36.

Wang, G. Z., Luo, H. Y., Wang, Y. R., Huang, H. Q., Shi, P.J., Yang, P. L., Meng, K., Bai, Y. G. & Yao, B. (2011). A novel cold-active xylanase gene from the environmental DNA of goat rumen contents: direct cloning, expression and enzyme characterization. *Biores. Technol*. **102(3)**, 3330-3336.

Wang, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. (2014). MetaCluster-TA: Taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics* **15(S1)**, S12.

Wang, Y., Hu, H. & Li, X. (2016). rRNAFilter: A fast approach for ribosomal RNA read removal without a reference database. *J. Comput. Biol.* **24(4)**, 368-375.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., *et al*. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560-565.

Warner, A. C. I. (1981). Rate of digesta passage through the gut of mammals and birds. *Nutrition Abstracts and Reviews* **Series B**, 789-820.

Wexler, M., Bond, P. L., Richardson, D. J. & Johnston, A. W. B. (2005). A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environ. Microbiol.* **7(12)**, 1917-1926.

Whistler, R. L. & Bemiller, J. N. (1958). Alkaline degradation of polysaccharides. *Adv. Carbohydr. Chem*. **13**, 289-329.

Wilson, D. B. (2008). Three microbial strategies for plant cell wall degradation. In *Incredible anaerobes: From physiology to genomics to fuels* (pp 289-297). Wiegel J., Maier, R. & Adams, M. (eds). Wiley-Blackwell, Oxford.

Wilson, D. B. (2009). Evidence for a novel mechanism of microbial cellulose degradation. *Cellulose* **16**, 723-727.

Wilson, C. A. & Wood, T. M. (1992a). Studies on the cellulase of the rumen anaerobic fungus *Neocallimastix frontalis*, with special reference to the capacity of the enzyme to degrade crystalline cellulose. *Enzyme Microbial Technol.* **14**, 258-264.

Wilson, C. A. & Wood, T. M. (1992b). The anaerobic fungus Neocallimastix frontalis – isolation and properties of a cellulosome-type enzyme fraction with the capacity to solubilize hydrogen-bond-ordered cellulose. *Appl. Microbiol. Biotechnol.* **37**, 125-129.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* **51**, 221-271.

Woese, C. R. (1996). Phylogenetic trees: Whither microbiology? *Curr. Biol.* **6**, 1060-1063.

Wojciechowicz, M. A. R. I. A., & Ziołecki, A. (1979). Pectinolytic enzymes of large rumen treponemes. *Appl. Environ. Microbiol.* **37(1),** 136-142.

Wolfenden, R., Snider, M., Redgway, C. & Miller, B. (1999). The temperature dependence of enzyme rate enhancements. *J. Am. Chem. Soc*. **121**, 7419-7420.

Wood, T. M., Wilson, C. A., McCrae, S. I. & Joblin, K. N. (1986). A highly-active extracellular cellulase from the anaerobic rumen fungus *Neocallimastix frontalis*. *FEMS Microbiol. Lett.* **34**, 37-40.

Wood, T. M. (1988). Preparation of crystalline, amorphous, and dyed cellulose substrates. *Methods in Enzymol.* **160**, 19-25.

Wood, D. E. & Salzburg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R36.

Wright, A. D., Dehority, B. A. & Lynn, D. H. (1997). Phylogeny of the rumen ciliates Entodinium, Epidinium and Polyplastron (Litostomatea:Entodiniomorphida) inferred from small subunit ribosomal RNA sequences. *J. Eukaryot. Microbiol*. **44(1)**, 61-67.

Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. (2011). WebMGA: A customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**, 444.

Wyman, C. E., Dale, B. E., Elander, R. T., Holtzapple, M., Ladisch, M. R., Lee, Y. Y., Mitchinson, C. & Saddler, J. N. (2009). Comparative sugar recovery and fermentation data following pretreatment of poplar wood by leading technologies. *Biotechnol. Prog*. **25**, 333-339.

Xia, Y., Ju, F., Fang, H. H. P. & Zhang, T. (2013). Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by metagenomics. *Plos One* **8(1)**, e53779.

Xie, G., Chain, P. S., Lo, C. C., Liu, K. L., Gans, J., Merritt, J. & Qi, F. (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular Oral Microbiology* **25**, 391-405.

Xiong, X., Frank, D. N., Robertson, C. E., Hung, S. S., Markle, J., Canty, A. J., McCoy, K. D., Macpherson, A. J., Poussier, P., Danska, J. S. & Parkinson, J. (2012). Generation and analysis of a mouse intestinal metatranscriptome through Illumina based RNA-sequencing. *Plos One* **7(4)**, e36009.

Yadav, S., Kundu, S., Ghosh, S. K. & Maitra, S. S. (2015). Molecular analysis of methanogen richness in landfill and marshland targeting 16S rDNA sequences. *Archaea* **13**;2015:563414.

Ying, J.-Y., Zhang, L.-M., Wei, W.-X. & He, J.-Z. (2013). Effects of land utilization patterns on soil microbial communities in an acid red soil based on DNA and PLFA analyses. *J. Soils Sediments* **13**, 1223-1231.

Yoder, R. D., Trenkle, A. & Burroughs, W. (1966). Influence of rumen protozoa and bacteria upon cellulose digestion. *J. Anim. Sci.* **25(3)**, 609-612.

Youssef, N. H., Couger, M. B., Struchtemeyer, C. G., Liggenstoffer, A. S., Prade, R. A., Najar, F. Z., Atiyeh, H. K., Wilkins, M. R. & Elshahed, M. S. (2013). Genome of the anaerobic fungus *Orpinomyces* sp. C1A revelas the unique evolutionary history of a remarkable plant biomass degrader. *Appl. Environ. Microbiol.* **79(15)**, 4620-4634.

Zalucki, Y. M. & Jennings, M. P. (2007). Experimental confirmation of a key role for non-optimal codons in protein export. *Biochem. Biophys. Res. Commun*. **355(1)**, 143-148.

Zalucki, Y. M., Beacham, I. R. & Jennings, M. P. (2009). Biased codon usage in signal peptides: A role in protein export. *Trends Microbiol*. **17(4)**, 146-150.

Zifcakova, L. & Baldrian, P. (2012). Fungal polysaccharide monooxygenases: New players in the decomposition of cellulose. *Fungal Ecol.* **5**, 481-489.

Zhang, Y.P. & Lynd, L. R. (2004). Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems. *Biotechnol. Bioeng.* **88(7)**, 797-824.

Zhang, D., Lax, A. R., Bland, J. M. & Allen, A. B. (2011). Characterisation of a new endogenous endo-b-1, 4-glucanase of Formosan subterranean termite (*Coptotermes formosanus*). *Insect Biochem. Mol. Biol.* **41(4)**, 211-218.

Zhilina, T. N. & Zavarzin, G. A. (1994). Alkaliphilic anaerobic communities at pH 10. *Curr. Microbiol.* **29**, 109-112.

Zhu, W., Lomsadze, A. & Borodovsky, M. (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res*. **38(12)**, e132.

Zhu, J. Y., Zhu, W., O'Bryan, P., Dien, B. S., Tian, S., Gleisner, R. & Pan, X. J. (2010). Ethanol production from SPORL pretreated lodgepole pine: Preliminary evaluation of mass balance and process energy efficiency. *Appl. Microbiol. Biotechnol*. **86**, 1355-1365.

Zumsteg, A., Luster, J., Goeransson, H., Smittenberg, R. H., Brunner, I., Bernasconi, S. M., Zeyer, J. & Frey, B. (2012). Bacterial, archaeal and fungal succession in the forefield of a receding glacier. *Microb. Ecol.* **63**, 552-564.

Zvereva, E. A., Fedorova, T. V., Kevbrin, V. V., Zhilina, T. N. & Rabinovich, M. L. (2006). Cellulase activity of a haloalkaliphilic anaerobic bacterium, strain Z-7026. *Extremophiles* **10**, 53-60.

Appendix A

DNA sequences of the ORFs determined to contain GH domains following sequencing and assembly of the P2E2 lake-fosmid library clone that had produced a positive Congo red assay reaction:

- **P2E21 (1,122 bp)-** GH10 (E-value $5.2e^{-84}$)

5’

ATGGTATACTTAACGCCCTACGATTGGAGGAATGGCATGAATCACGCACGCATACCCGTTCT

GGTTGCCCTTTGCGCGGCAATCATCGTTTCCGTCCCGGCATGCGGGGCAAAAAAGACCCTTC

GCGCGAACGCGGACGAGCGAAAGCTTTCCTTTGGCGTTTCAGTGCAGGCGGGAGACGTCTT

TGACCCCGTAGCGATCAAACTCATACAGAGTAATTTTAACGAGATCGTACCCGAAAATACCA

TGAAATGGCGAAATATCAGACCGACGAAGGGCTTCTGGAACTGGTCCGACATGGACGGCA

TGGTCGCGTTCGCGGAAAAGAACAAAATCCGGATGAAGGGACACGTCTTCCTGTGGCATCA

GCAAAACCCGCCCTACGTCGACGGTCTTAAAACCCGTGACGAGGCGCTCGCCCTCATGATC

GAACAGATCTCTACCGTTATGGCGCGATATAAAGGAAGAATCTACGAATACGATGTCGCGA

ACGAGGTTCTCAACGACGACGGCTCGATGCGGGGTACGGTCTGGTACCGGACGATCGGCG

CCGACTATCTCGACATCGCTTTTCGCACCGCGCACTCGGCCGACCCCGCTGCGCGCCTCATC

TTAAACGATTACGGAAACGAGTACGCGGGAACGCCGAAGGGCGACGCCTTCTACGAACTC

GTCAAGGGACTCGTAGCCCGCGGTGTGCCGATCTCGGGCGTCGGGCTCCAGCTCCACGTGC

AAGCGCACGATATCGTCAACGAGAGCGCCCTCCGCGAAACGATCAAGCGCTTCCGCGACCT

TGGGCTTTTCGTCTCGTTCACCGAGATCGACGTTCGCGTCGCGATGCCCGTCACTTCAGAAA

AAGAGGCCGAGCAGGTCGCCGCCTACACCAAGCTCATGGAAGTCGCCTTAAGCGAACCCA

ATGCTGCGAGCTTTATCATGTGGGGCTATACCGACAAGCTCAGCTGGATCCCTGGGTTTTTC

CCCGGTTACGGATCCGCGCACCTGTTTGACCGCGAGGTAAAACCGAAGGCGGCGTATCGG

GCGATCGAGGCCATGGTCGCTCAACAGGAAAAAGGGAAACAGCGGGCGGCAAGCGTCGC

AGGAACCGCAAAAAAGAAAGCGTAG

3’

- **<u>P2E22</u> (1,221 bp)-** GH5 (E-value 6.4e^{-60})

5’

ATGAGTGCTTACTCTTCATTATTGAGACGATGCTCCGTATTCTTTGCCCTTCTTTTCGCGACTG

TTCTGACTGCCTGCGCGTCAGGCTCGAGCGATACCCTCTCGGCCAACCTTCCCGTGGAGCCG

GGGTATACCTCCGCAGCGATAGCGCCCGAGACCGCGGATTTTTCGAAGAGCGCCATCGCGT

TCGTTGCCGGAATGGGGACTGGTTGGAACCTCGGAAACACCCTGGACGCCACCGGAACGG

CTGACCTGACCTGCGAAACCGCCTGGGGACAGCCCAAGACTACCAAGGCGATGATTGCCG

GACTCAAGGCCTCGGGCATCAAGACGATCCGTATCCCCATATCCTGGCATGATCACGTCGAC

TCGGCGTTTACCGTGGATAGTGCGTGGATGGCACGCGTGAAGGAAGTCGTTGACTACGCG

ATCGACGAGGGCCTGTACGTCATCATCAACATCCACCATGACAACGATAAGGCGTATTATTT

CCCCGACAGGGCCCATGCGGATCGCTCGCGCGAGTACGTCAGGCGCGTATGGAAGCAGGT

CGCCCTCGAGTTTCGCAACTACGACGAGCACCTCATCTTCGAGATCCTGAACGAGCCTCGCC

TCGTCGGCTCTGCGAACGAATGGAATTGGAGCGACGCGGACTCGAGTCTCGTCGCGGCGG

CGAGCGTCATCGTCGATCTCGAGCAAAGCGCGCTCACCGCGATCCGCACGACGGGAAGCA

ACAACGAGTACCGGTACGTCATGATCACGCCCTACGTCGCGTCTCCTTGGGCCGCCCTATCG

CCAAAGTTCAGCATCCCGAGCGACACCGCGACCGATAAGCTCATCCTATCCGTGCATGCCTA

TCAACCGTACAGCTTCGCGATGCAGGACCCGGGAGAGACGAAGTTTACCGCGTCTCATAAA

ATCGAGATCGATACCTTCATGGGTAACCTCAATACGAAGTTCGTTCAGGGAAAGGGTATGC

CGGTAATCATCGGGGAATACGGCGCGACGAACAAGAACAACCTCGCCGAGCGCGTCGCGT

GGTTTTCATATTACGTGGGTAAGGCGAAATCCTACGGCATGGTGACCGTCCTCTGGGACAA

CGGGAACTCACAGGTGCCAAGCTCGGGGAAGTTTAGCGAGCTCTACGGCTTCTATAACCGC

ACCGCGCAAACCTGGTACTTCCCCACGATCCTTCAGGCAATTATCGACGCGTCGAAATAG

3’